Music Recommender System

Student Name: Kerri Humphrey (khumphrey@ryerson.ca)
Student ID#: 500786829

Introduction

With the numerous music streaming and retail services and the growing number of artists and songs, the need exists to be able to find new and desirable content based on preferences and how users engage with music. Based on individuals' attitudes, preferences and demographics, combined with how songs and artists are described, can predictions be made about how much individuals will like a particular artist or song. The objective of this project is to develop a recommender system combining these various attributes using a two-step approach, first clustering users based on demographic characteristics and preferences and content-based filtering to select artists similar to those they have previously rated highly. The dataset for this project is the EMI One Million Interview Dataset, utilizing recommender system algorithms through R and Spark.

Literature Review

Recommender systems vary in complexity from the basic type that make recommendations based on the popularity of items, i.e. the most purchased, most listened to or highest rated items are recommend to any and everyone, to more complex systems that take into consideration item features and user characteristics and preferences in order to provide more personalized recommendations. Recommendation systems are used to recommend a variety of different types of items from music, research papers and books to consumer products. Furthermore, these systems can make use of explicit information such as users ratings of items or implicit information such as their behaviour (e.g. purchase behaviour or monitoring their behaviour in other services) and apply predictive analytics to make personalized recommendations to users in order to create the most positive experience.

While numerous algorithms and approaches have been developed to recommend items y users, they can be categorized into three major categories: Collaborative Filtering, Content-based Filtering and a Hybrid Approach which combines the two techniques. A number of research papers have been written evaluating the merits of each approach.

Collaborative Filtering

Collaborative Filtering is a process of filtering or evaluating items which relies on based on the opinions, ratings, purchases, behaviours or other past interactions with items of others. This approach has the advantage of known popularity of items and is based on the premise that if a lot of users like it, there is a good likelihood that others will too (Schafer et al, n.d.). However, Verma, Patel and Patel (2015) note that this approach does not take into account user preferences or item characteristics. Furthermore, McFee et al. (2012) note that with this approach there is no basis on which to determine if users will like a new item, since new items cannot be recommended until they are purchased, listened to, rated or

otherwise reacted to thus making it difficult for new music to be explored or discovered. Another disadvantage of this approach can occur if there are more items than users (Gipp, Beel & Hentschel, 2009).

Content-Based Filtering

Content-based Filtering on the other hand is based on characteristics or features of items and users' preferences. These approaches recommend additional items that are similar to those that users have purchased or liked in the past. One advantage identified of Content-based recommender systems highlighted by Narayanan and Cherukuri (2014) is that they are not prone to the cold start problem encountered when faced with new users or new items provided they possess the appropriate descriptive content. However, a disadvantage is that the descriptive content/information of the users and items needs to be maintained which could be costly.

Hybrid Approach

Many researchers propose combining the two methods in what is commonly referred to as Hybrid Recommenders in an effort to improve performance. Narayanan and Cherukuri (2014) note that combining the two approaches can address the disadvantages of each. Researchers have proposed a two-stage process for recommender systems combining the two-filtering methods. Verma et al. (2015) suggest two ways in which the Hybrid Filtering Approach can be applied. The first is by applying Content-based Filtering and Collaborative Filtering separately and then combining the results, while the second approach involves first applying collaborative filtering and then applying content-based filtering on the results. Verma et al. (2015) develop a Hadoop based recommendation system applying a Hybrid Filtering Approach applied to user reviews, opinions, remarks comments and complaints combined with ratings, ranks, content and reviewers' behaviour

Another example is McFee et al. (2012) who developed a recommender system for music by learning the content similarity. They used a content based similarity method initially and then collaborative similarity method is imposed on the results which avoids the cold start problem. McFee et al. (2012) note that the most successful approaches applied to a variety of recommendation tasks is collaborative filtering which relies on "the wisdom of crowds" to infer similarities between items and recommend new items to users by representing and comparing these items in terms of the people who use them. While they note that collaborative filtering forms the basis for many state-of-the-art recommendation systems, it is incapable of recommending items that have not yet been consumed or rated. However, they propose making use of content-based similarities in an effort to alleviate the cold-start problem, allowing music recommendation to be extended to new or lesser known songs.

Similarly, Zhang et al. (2010) also present a two stage recommendation algorithm based on K-means clustering which they found had higher accuracy in the context of mobile e-commerce recommendations. In their study they cluster users based on profile information to find neighbour users, than apply collaborative filtering to identify items to recommend.

Another study by Kavitha and Mohanapriya (2015) also supports a two step process to recommendation, starting with the clustering of users based on a similarity matrix using profile

information before applying a collaborative filtering technique of tensor factorization to refine the clustering results and provide recommendations. Vinodhini et al. (2014) developed a recommender system for books that analyzes the interests of the users and features of the book in order to increase the accuracy of the recommendations. In their approach, users are clustered based their profile information using K-means clustering algorithms and this combined with ratings lists are used to provide recommendations.

Dataset

The dataset for this project is a subset of the EMI One Million Interview Dataset from a 2012 Kaggle Competition (EMI Music Data Science Hackathon) which consists of three data files:

1)***Users:*** User profile information consisting of gender, age, employment status, respondents view on the importance of music in their life, number of daily hours spent listening to music of their choice, number of daily hours spent listening to music they have not chosen and 19 questions rated from 0 to 100 about music habits/attitudes.

2) ***Words:*** 82 words users describe artist artists, familiarity with artists, whether they own music by the artist, and a rating from 0 to 100 indicating the extent they like or dislike listening to the artist.
Forty- three of the descriptor words will be excluded due to the high proportion missing values. The 40 descriptor words that will be used are: Aggressive, Edgy, Thoughtful, Serious, Good Lyrics, Unattractive, Confident, Youthful, Boring, Current, Stylish, Cheap, Calm, Outgoing, Inspiring, Beautiful, Fun, Authentic, Credible, Cool, Catchy, Timeless, Depressing, Original, Talented, Distinctive, Approachable, Trend Setter, Noisy, Upbeat, Energetic, None of these, Sexy, Fake, Cheesy, Unoriginal, Dated, Playful, Arrogant, Warm. Additionally, Own Artists Music and Like Artist fields will be excluded due to proportion of missing values.

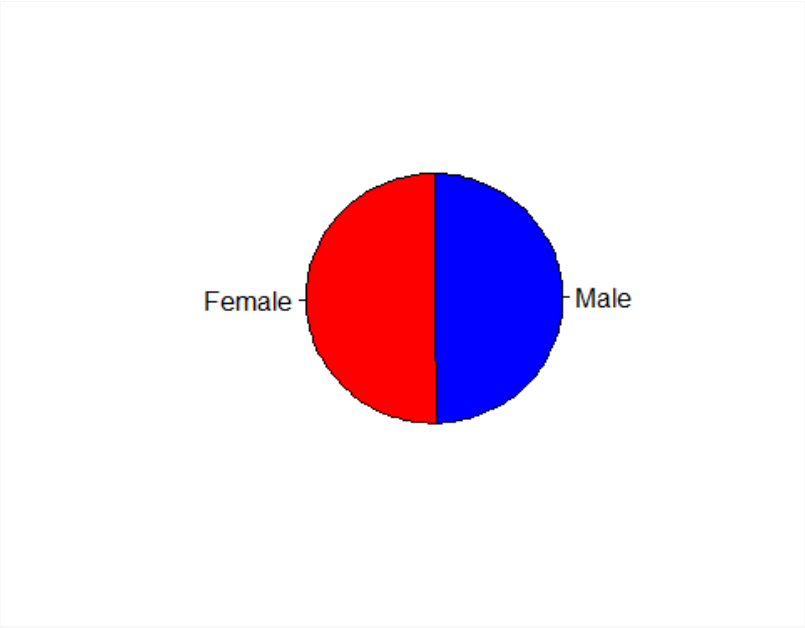
3)***Data:*** Dataset of ratings from 0 to 100 by users of a number of songs/tracks

Data Statistics

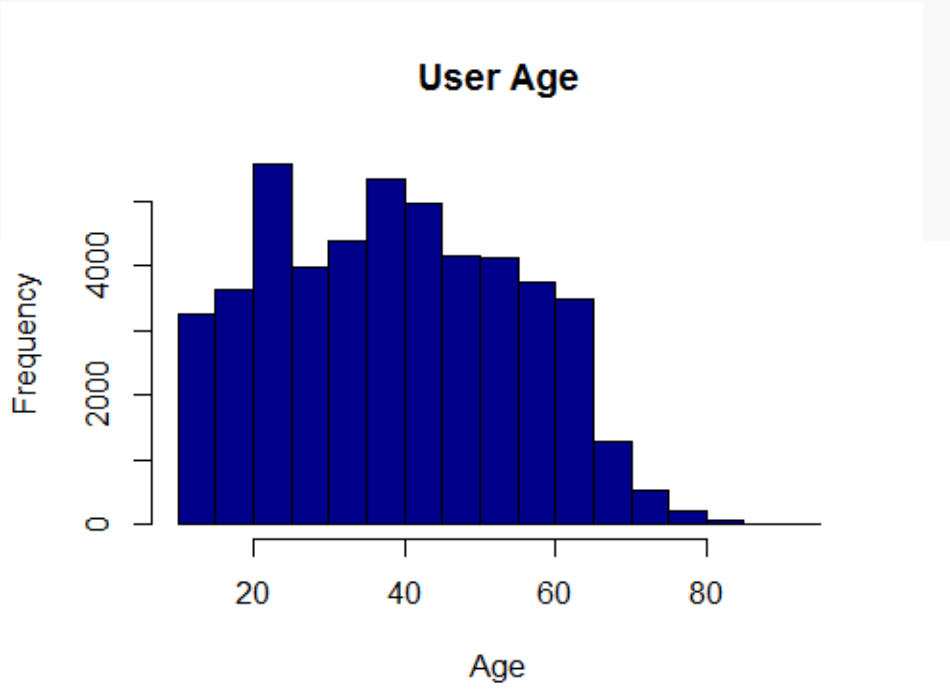
- 50 Artists
- 184 Tracks
- 49,479 Users
- 188,690 Ratings
    - Avg. Rating = 36.44
    - Average Number of ratings per user = 3.81
    - Average Number of Ratings per Artist = 3,773.8
- 118,301 ratings of descriptors of artists (words)

Descriptive Statistics of Users Dataset

```
##     GENDER
##  Female:24503
##  Male  :24142
```

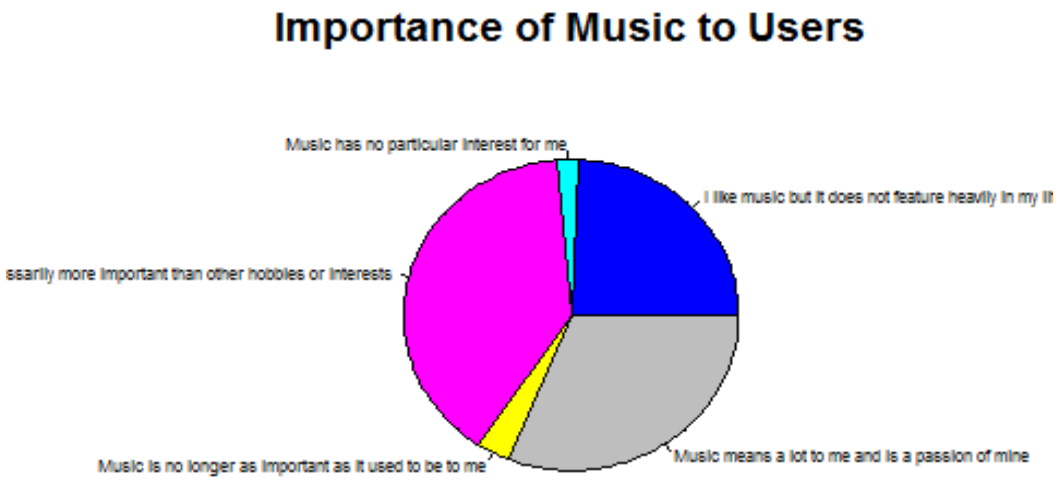



```
##       AGE
##  Min.   :13.00
##  1st Qu.:25.00
##  Median :39.00
##  Mean   :39.28
##  3rd Qu.:52.00
##  Max.   :94.00
```



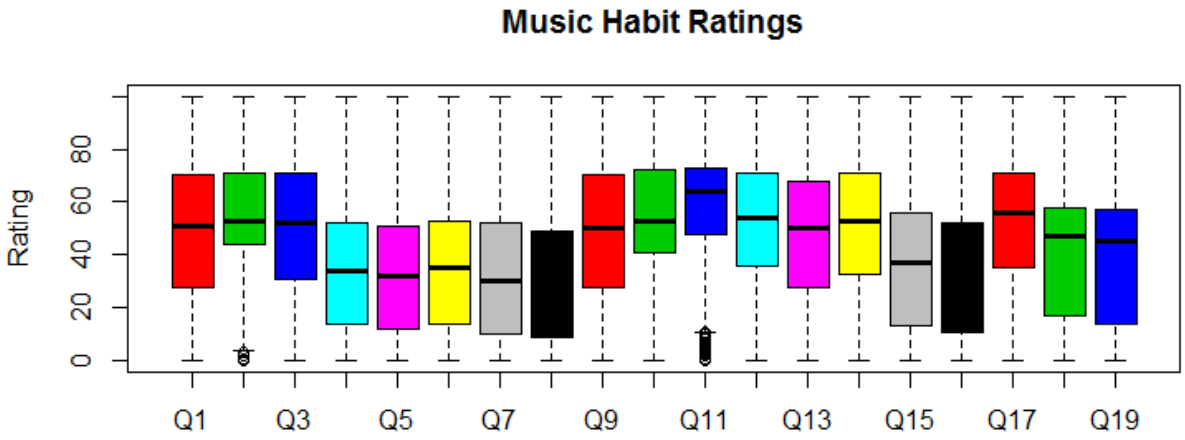
```
      Importance of Music

## I like music but it does not feature heavily in my life       11790
##
## Music has no particular interest for me                        1037
##
## Music is important to me but not necessarily more important
   than other hobbies or interests                                19132
##
## Music is no longer as important as it used to be to me         1604
##
## Music means a lot to me and is a passion of mine               15082
##
```



Attitudes toward the importance of music to users showed statistically significant differences in average artist track ratings and will therefore be used to build the model.

```
##                          WORKING
##  Employed 30+ hours a week                                :13617
##  Full-time student                                        : 5105
##  Employed 8-29 hours per week                             : 4086
##  Retired from full-time employment (30+ hours per week): 3292
##  Full-time housewife / househusband                       : 2627
##  (Other)                                                  : 6793
##  NA's                                                     :13125
##               REGION
##  Centre          : 2846
##  Midlands        :11844
##  North           :16741
##  North Ireland   :  138
##  Northern Ireland:  769
##  South           :15267
##  NA's            : 1040
##
```



Q1	Q2	Q3	Q4
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 28.00	1st Qu.: 44.00	1st Qu.: 31.00	1st Qu.: 14.00
Median : 51.00	Median : 53.00	Median : 52.00	Median : 34.00
Mean : 49.11	Mean : 54.62	Mean : 51.28	Mean : 37.31
3rd Qu.: 70.00	3rd Qu.: 71.00	3rd Qu.: 71.00	3rd Qu.: 52.00
Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00

Q5	Q6	Q7	Q8
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 12.00	1st Qu.: 14.00	1st Qu.: 10.00	1st Qu.: 9.00
Median : 32.00	Median : 35.00	Median : 30.00	Median : 23.00
Mean : 34.59	Mean : 39.33	Mean : 33.85	Mean : 29.16
3rd Qu.: 51.00	3rd Qu.: 53.00	3rd Qu.: 52.00	3rd Qu.: 49.00
Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00

Q9	Q10	Q11	Q12
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 28.00	1st Qu.: 41.00	1st Qu.: 48.00	1st Qu.: 36.00
Median : 50.00	Median : 53.00	Median : 64.00	Median : 54.00
Mean : 47.83	Mean : 55.01	Mean : 58.64	Mean : 53.67
3rd Qu.: 70.00	3rd Qu.: 72.00	3rd Qu.: 73.00	3rd Qu.: 71.00
Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00

Q13	Q14	Q15	Q16
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 28.00	1st Qu.: 33.00	1st Qu.: 13.00	1st Qu.: 11.00
Median : 50.00	Median : 53.00	Median : 37.00	Median : 32.00
Mean : 46.96	Mean : 53.45	Mean : 39.66	Mean : 35.58
3rd Qu.: 68.00	3rd Qu.: 71.00	3rd Qu.: 56.00	3rd Qu.: 52.00
Max. :100.00	Max. :100.00	Max. :100.00	Max. :100.00
			NA's :6435

```
      Q17              Q18              Q19           LIST_BACK2
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
 1st Qu.: 35.00   1st Qu.: 17.00   1st Qu.: 14.00   1st Qu.: 1.000
 Median : 56.00   Median : 47.00   Median : 45.00   Median : 2.000
 Mean   : 53.83   Mean   : 42.23   Mean   : 41.36   Mean   : 3.051
 3rd Qu.: 71.00   3rd Qu.: 58.00   3rd Qu.: 57.00   3rd Qu.: 4.000
 Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :24.000
                  NA's   :13125    NA's   :13125    NA's   :5825

   LIST_OWN2
 Min.   : 0.000
 1st Qu.: 1.000
 Median : 1.000
 Mean   : 2.349
 3rd Qu.: 3.000
 Max.   :24.000
 NA's   :5939
```

Descriptive Statistics of Words Dataset

```
##                         HEARD_OF
##  Never heard of                            :61892
##  Heard of                                  :22878
##  Heard of and listened to music EVER       :19914
##  Heard of and listened to music RECENTLY:12577
##  Ever heard music by                       :  579
##  (Other)                                   :  437
##  NA's                                      :   24


##        OWN_ARTIST_MUSIC                    LIKE_ARTIST     Uninspired
##  Don't know                      :  820   Min.   :  0.00   0   :23832
##  Own a little of their music     :11428   1st Qu.: 31.00   1   : 2322
##  Own a lot of their music        : 4298   Median : 49.00   NA's:92147
##  Own all or most of their music: 1535   Mean   : 48.12
##  Own none of their music         :15426   3rd Qu.: 65.00
##  NA's                            :84794   Max.   :100.00
##                                           NA's   :84993


##  Sophisticated Aggressive   Edgy       Sociable     Laid.back
##  0   :19507    0   :92391   0:107263   0   :19296   0   :18152
##  1   : 1217    1   : 5186   1: 11038   1   : 1428   1   : 2572
##  NA's:97577    NA's:20724              NA's:97577   NA's:97577
##
##
##  Wholesome     Uplifting    Intriguing   Legendary       Free
##  0   :  1002   0   :18795   0   :19367   0   :   710   0   :16480
##  1   :    38   1   : 1929   1   : 1357   1   :   330   1   : 4244
##  NA's:117261   NA's:97577   NA's:97577   NA's:117261   NA's:97577
##
```

```
##  Thoughtful Outspoken    Serious      Good.lyrics  Unattractive
##  0:107676   0   :19830   0   :92696   0   :99964   0   :90913
##  1: 10625   1   :  894   1   : 4881   1   :18337   1   : 6664
##             NA's:97577   NA's:20724                NA's:20724
##
##
##  Confident      Old          Youthful       Boring      Current
##  0   :82899   0   :   893   0   :105349   0   :71727   0:100092
##  1   :14678   1   :   147   1   : 11912   1   :15353   1: 18209
##  NA's:20724   NA's:117261   NA's:  1040   NA's:31221
##
##
##  Colourful    Stylish     Cheap       Irrelevant   Heartfelt
##  0   :18365   0:107405   0   :93871   0   :24938   0   :18436
##  1   : 2359   1: 10896   1   : 3706   1   : 1216   1   : 2288
##  NA's:97577              NA's:20724   NA's:92147   NA's:97577


##    Calm       Pioneer       Outgoing     Inspiring    Beautiful
##  0   :82851   0   :   913   0   :92848   0   :90882   0:105656
##  1   :14726   1   :   127   1   : 4729   1   : 6695   1: 12645
##  NA's:20724   NA's:117261   NA's:20724   NA's:20724
##
##
##  Fun        Authentic  Credible   Way.out      Cool        Catchy
##  0:106113   0:104083   0:106963   0   :20211   0:101621   0   :93023
##  1: 12188   1: 14218   1: 11338   1   :  513   1: 16680   1   :24238
##                                   NA's:97577              NA's: 1040
##
##
##  Sensitive    Mainstream   Superficial  Annoying       Dark
##  0   :90383   0   :43248   0   :93173   0   :21747   0   :   830
##  1   : 7194   1   : 3006   1   : 4404   1   : 4407   1   :   210
##  NA's:20724   NA's:72047   NA's:20724   NA's:92147   NA's:117261
##
##
##  Passionate Not.authentic Background   Timeless      Approachable
##  0:109046   0   :25252    0   :19228   0:108055      0:107796
##  1:  9255   1   :  902    1   : 1496   1: 10246      1: 10505
##             NA's:92147    NA's:97577
##
##  Depressing   Original   Talented   Worldly       Distinctive
##  0   :89227   0:101241   0:100334   0   :   948   0:95630
##  1   : 8350   1: 17060   1: 17967   1   :    92   1:22671
##  NA's:20724                         NA's:117261
##
##   Genius      Trendsetter  Noisy        Upbeat       Relatable
##  0   :19754   0:113553    0   :90183   0   :104759   0   :45074
##  1   :  970   1:  4748    1   : 7394   1   : 12502   1   : 1180
##  NA's:97577               NA's:20724   NA's:  1040   NA's:72047
```

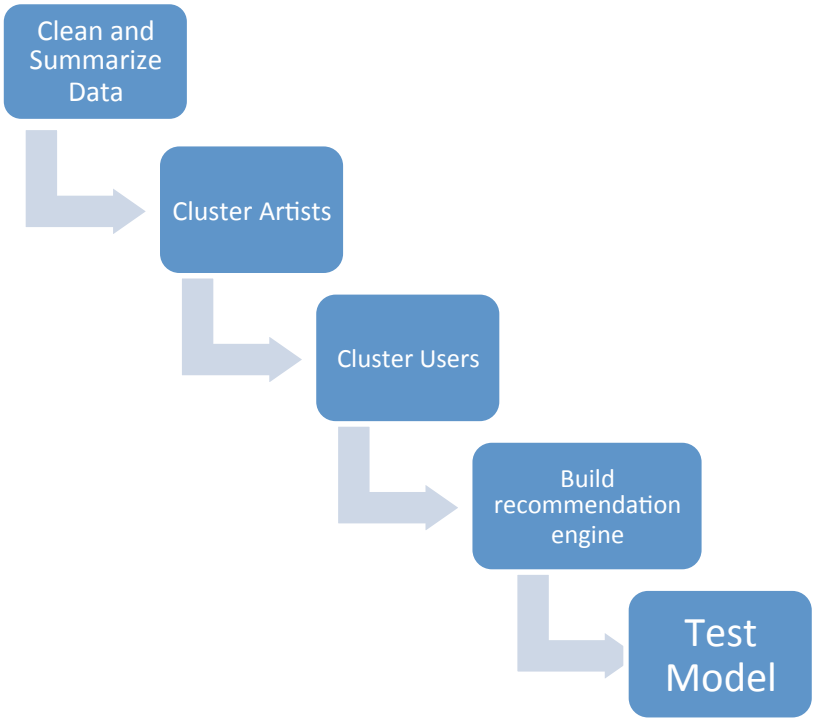
```
##  Energetic  Exciting     Emotional    Nostalgic     None.of.these
##  0:102368   0   :18796   0   :18117   0   :   817   0:107803
##  1: 15933   1   : 1928   1   : 2607   1   :   223   1: 10498
##             NA's:97577   NA's:97577   NA's:117261
##
##
##  Progressive   Sexy         Over       Rebellious     Fake
##  0   :   866   0:113785   0   :87971   0   :19302   0   :94424
##  1   :   174   1:  4516   1   : 2186   1   : 1422   1   : 3153
##  NA's:117261              NA's:28144   NA's:97577   NA's:20724
##
##   Cheesy      Popular      Superstar    Relaxed      Intrusive
##  0   :91821   0   :16475   0   :45512   0   :18620   0   :25587
##  1   : 5756   1   : 3209   1   :  742   1   : 2104   1   :  567
##  NA's:20724   NA's:98617   NA's:72047   NA's:97577   NA's:92147
##  Unoriginal    Dated         Iconic       Unapproachable Classic
##  0   :89418   0   :109329   0   :   774   0   :95664     0   :94667
##  1   : 8159   1   :  7932   1   :   266   1   : 1913     1   :10568
##  NA's:20724   NA's:  1040   NA's:117261   NA's:20724     NA's:13066
##
##  Playful      Arrogant     Warm       Soulful
##  0   :91856   0   :94013   0:107413   0   :17822
##  1   : 5721   1   : 3564   1: 10888   1   : 1862
##  NA's:20724   NA's:20724              NA's:98617
```

Correlation of Numeric Variables with Users' Ratings of Songs

```
      AGE         Q1         Q2         Q3         Q4         Q5          Q6
-0.04435332 0.1482717 0.1417984 0.1488157 0.09707373 0.05462955 -0.02521545
   Q7         Q8          Q9        Q10        Q11        Q12       Q13       Q14
0.09395977 0.0923180 -0.0640856 0.1182688 0.1894556 0.17539 0.146904 0.148165
    Q15        Q16        Q17        Q18        Q19   LIST_BACK2  LIST_OWN2
0.1352577 0.1540842 0.1874442 0.1588649 0.1420988 0.06086321 0.04961939
```

The numeric variables in the dataset show weak but statistically significant correlations with the artist track ratings and will therefore be used in building the model.

Approach



Step 1: Clean and Summarize Data

Convert Artist, Track and User IDs to categorical variables from numeric. Fill in missing values for age with average age by gender. Merge text and numeric responses or List_Own and List_Back variables and convert to numeric variables. Obtain descriptive and inferential statistics to identify variables to be considered to build model.

https://github.com/Special-Keh/CapstoneProject

Step 2: Cluster Users

Using the profile information in the Users data, cluster users using K-Means Clustering based on their demographic characteristics and attitudes toward music to identify similar users to identify highly rated tracks/artists to recommend.

Step 3: Cluster Artists

Using the Words data cluster Artists based on the words respondents use to describe them using

Step 4: Develop Recommender System

Develop regression model to predict if a user will like a song/artist. Using the profile information in the Users data, cluster users based on their demographic characteristics to identify similar users to identify highly rated tracks/artists to recommend.

Step 5: Test Model

Evaluate the effectiveness of the model on the test dataset

Results

Four regression models were tested on the data to determine the best approach to predict ratings of artists in order to recommend new artists to users.

Model 1 - Regression Model on Artist Descriptors Data

Binary coded descriptor variables and users' level of familiarity with the artist were used to predict average artist ratings given by a particular user.

```
Training Model Summary Statistics
       Root Mean Squared Error (RMSE): 15.249278087801935
       Mean Squared Error (MSE): 232.5404821991162
       Mean Absolute Error (MAE): 12.005791790072688
       Explained variance = 223.99115402100944
       r2: 0.5097573558843733
```

The regression model using artist descriptors explains approximately 51% of the explained variance in average artist ratings.

Model 2 - Regression Model on User Profile Data

User data consisting of demographic information and preferences about the way users interact with music were used to predict users' ratings of various tracks from a variety of artists.

```
Training Model Summary Statistics
       Root Mean Squared Error (RMSE): 21.849239705598293
       Mean Squared Error (MSE): 477.389275712693
       Mean Absolute Error (MAE): 17.80784203098645
       Explained variance = 29.318781401853883
       r2: 0.0668802788275743
```

The second regression model based on artist preferences did not perform as well in predicting users' ratings of music tracks. The evaluation metrics of this model indicate that the error/variance between the predicted ratings and the actual ratings was higher than that for Model 1. This model explained on 6% of the observed variance (R^2).

Model 3 - Joined Dataset of Artist and User Data Predicting Average Artist Ratings

Combined user data and artist descriptors used to predict average ratings of a number of tracks for a given artist.

```
Training Model Summary Statistics
       Root Mean Squared Error (RMSE): 14.613401689139478
       Mean Squared Error (MSE): 213.55150892814453
       Mean Absolute Error (MAE) = 11.491453396785442
       Explained variance = 215.21338120059315
       r2: 0.5216866265828447
```

The third model with user profile and music preference information combined with artist descriptors produced a model that explained a slightly higher proportion of the variance in average user ratings (R^2 = 52%). The evaluation metrics indicate that the error/variance between the predicted and actual ratings is lower than both previous models.

Model 4 - Joined Dataset of Artist and User Data Predicting Artist Track Rating

The fourth model, similar to the third model, uses user information and artist descriptions to predict the ratings of individual tracks.

```
Training Model Summary Statistics
      Root Mean Squared Error (RMSE): 16.710629204611397
      Mean Squared Error (MSE): 279.2451284140113
      Mean Absolute Error (MAE) = 13.244935369659505
      Explained variance = 213.40552425851715
      r2: 0.4538742311159145
```

The fourth model, ranked third in terms of performance.

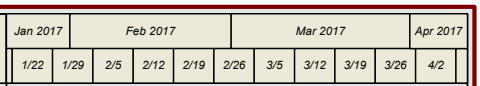
Based on the results, best performance occurs in predicting the average rating that a user would give to an artist based on various tracks rather than predicting the ratings of a particular track. The combined dataset of Artist Descriptors and User Profile Details to predict average artist ratings provided the best model of predicting how much users would like various artists. The evaluation metrics of the various models indicate that artist descriptors are more effective in predicting user ratings than user attitudes toward music. This model would allow the identification of new, unrated artists for recommendation to users.

Link to code:

https://github.com/Special-Keh/CapstoneProject/blob/master/Capstone%20Project_Mar17.scala

Project Timeline

ID	Task Name	Start	Finish	Duration
1	Review of literature and projects on the topic and identify data set	1/18/2017	1/30/2017	9d
2	Clean, code and transform data	1/30/2017	2/20/2017	16d
3	Test predictive algorithms and data models	2/20/2017	3/20/2017	21d
4	Build Recommender framework/ interface	3/20/2017	4/11/2017	17d



References

Gipp, B., Beel, J., & Hentschel, C. (2009). “Scienstein: A Research Paper Recommender System,” in Proceedings of the *International Conference on Emerging Trends in Computing* (ICETiC’09), Virudhunagar, India.

Kavitha, M & Mohanapriya, A. (Sept. 2015). "A Hybrid Cluster Based Collaborative Filtering with Tensor Factorization Approach for Recommendation System in Big Data". *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 3, Issue 9, pp. 8711-8719

McFee, B., Barrington, L., & Lanckriet, G. (Oct. 2012). "Learning Content Similarity for Music Recommendation". *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No 8, pp. 2207-2218.

Narayanan, M., & Cherukuri, A. K. (2016). "A study and analysis of recommendation systems for location based social network (LBSN) with big data". *IIMB Management Review*, 28, pp. 25 - 30.

Schafer, J.B., Frankowski, D., Herlocker, J. & Sen, S. (2006). *Collaborative Filtering Recommender System* http://faculty.chas.uni.edu/~schafer/publications/CF_AdaptiveWeb_2006.pdf

Verma, J. P., Patel, B. & Patel, A. (April 2015). "Big Data Analysis: Recommendation System with Hadoop Framework". *IEEE International Conference on Computational Intelligence & Communication Technology*, pp. 92-97.

Vinodhini, S. Rajalakshmi, V., & Govindarajalu, B. (April 2014). "Building Personalised Recommendation System With Big Data and Hadoop Mapreduce". *International Journal of Engineering Research & Technology (IJERT)*, Vol. 3, Issue 4, pp. 2310 - 2316.

Zhang, F., Liu, H., & Chao, J. (2010). “A Two-stage Recommendation Algorithm Based on K-means Clustering In Mobile E-commerce”, *Journal of Computational Information Systems*, Vol. 6, Issue 10, pp. 3327-3334.