# 1990-2010 Song Popularity: The Data Dictionary

*Saeid Abolfazli*

*May 9, 2016*

## Introduction

The music industry has a well-developed market with a global annual revenue around $15 billion. The recording industry is highly competitive and is dominated by three big production companies which make up nearly 82% of the total annual album sales.

Artists are at the core of the music industry and record labels provide them with the necessary resources to sell their music on a large scale. A record label incurs numerous costs (studio recording, marketing, distribution, and touring) in exchange for a percentage of the profits from album sales, singles and concert tickets.

Unfortunately, the success of an artist's release is highly uncertain: a single may be extremely popular, resulting in widespread radio play and digital downloads, while another single may turn out quite unpopular, and therefore unprofitable.

Knowing the competitive nature of the recording industry, record labels face the fundamental decision problem of which musical releases to support to maximize their financial success.

## Prediction Goal?

How can we use analytics to predict the popularity of a song? In this assignment, we challenge ourselves to predict whether a song will reach a spot in the Top 10 of the Billboard Hot 100 Chart.

Taking an analytics approach, we aim to use information about a song's properties to predict its popularity.

## Dataset

The dataset `songs.csv` consists of all songs which made it to the Top 10 of the *Billboard Hot 100* Chart from *1990-2010* plus a sample of additional songs that didn't make the Top 10. This data comes from three sources: Wikipedia, Billboard.com, and EchoNest.

The variables included in the dataset either describe the artist or the song, or they are associated with the following song attributes: time signature, loudness, key, pitch, tempo, and timbre.

### Data Description

Here's a detailed description of the variables:

- **year** = the year the song was released
- **songtitle** = the title of the song
- **artistname** = the name of the artist of the song
- **songID and artistID** = identifying variables for the song and artist
- **timesignature and timesignature_confidence** = a variable estimating the time signature of the song, and the confidence in the estimate

- **loudness =** a continuous variable indicating the average amplitude of the audio in decibels
- **tempo and tempo_confidence =** a variable indicating the estimated beats per minute of the song, and the confidence in the estimate
- **key and key_confidence =** a variable with twelve levels indicating the estimated key of the song (C, C#, . . ., B), and the confidence in the estimate
- **energy =** a variable that represents the overall acoustic energy of the song, using a mix of features such as loudness pitch = **a continuous variable that indicates the pitch of the song
- **timbre_0_min, timbre_0_max, timbre_1_min, timbre_1_max, . . . , timbre_11_min, and timbre_11_max** = variables that indicate the minimum/maximum values over all segments for each of the twelve values in the timbre vector (resulting in 24 continuous variables)
- **Top10 =** a binary variable indicating whether or not the song made it to the Top 10 of the Billboard Hot 100 Chart (1 if it was in the top 10, and 0 if it was not)

```
file <- file.path("data","songs.csv")
songs <- read.csv(file)
str(songs)
```

```
## 'data.frame':    7574 obs. of  39 variables:
##  $ year                    : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ songtitle               : Factor w/ 7141 levels " l'or_e des bois""| __truncated__,..: 6204 5522 1
##  $ artistname              : Factor w/ 1032 levels "50 Cent","98 Degrees",..: 3 3 3 3 3 3 3 3 3 12 .
##  $ songID                  : Factor w/ 7549 levels "SOAACNI1315CD4AC42",..: 595 5439 5252 1716 3431 1
##  $ artistID                : Factor w/ 1047 levels "AR00B1I1187FB433EB",..: 671 671 671 671 671 671 6
##  $ timesignature           : int  3 4 4 4 4 4 4 4 4 4 ...
##  $ timesignature_confidence: num  0.853 1 1 1 0.788 1 0.968 0.861 0.622 0.938 ...
##  $ loudness                : num  -4.26 -4.05 -3.57 -3.81 -4.71 ...
##  $ tempo                   : num  91.5 140 160.5 97.5 140.1 ...
##  $ tempo_confidence        : num  0.953 0.921 0.489 0.794 0.286 0.347 0.273 0.83 0.018 0.929 ...
##  $ key                     : int  11 10 2 1 6 4 10 5 9 11 ...
##  $ key_confidence          : num  0.453 0.469 0.209 0.632 0.483 0.627 0.715 0.423 0.751 0.602 ...
##  $ energy                  : num  0.967 0.985 0.99 0.939 0.988 ...
##  $ pitch                   : num  0.024 0.025 0.026 0.013 0.063 0.038 0.026 0.033 0.027 0.004 ...
##  $ timbre_0_min            : num  0.002 0 0.003 0 0 ...
##  $ timbre_0_max            : num  57.3 57.4 57.4 57.8 56.9 ...
##  $ timbre_1_min            : num  -6.5 -37.4 -17.2 -32.1 -223.9 ...
##  $ timbre_1_max            : num  171 171 171 221 171 ...
##  $ timbre_2_min            : num  -81.7 -149.6 -72.9 -138.6 -147.2 ...
##  $ timbre_2_max            : num  95.1 180.3 157.9 173.4 166 ...
##  $ timbre_3_min            : num  -285 -380.1 -204 -73.5 -128.1 ...
##  $ timbre_3_max            : num  259 384 251 373 389 ...
##  $ timbre_4_min            : num  -40.4 -48.7 -66 -55.6 -43.9 ...
##  $ timbre_4_max            : num  73.6 100.4 152.1 119.2 99.3 ...
##  $ timbre_5_min            : num  -104.7 -87.3 -98.7 -77.5 -96.1 ...
##  $ timbre_5_max            : num  183.1 42.8 141.4 141.2 38.3 ...
##  $ timbre_6_min            : num  -88.8 -86.9 -88.9 -70.8 -110.8 ...
##  $ timbre_6_max            : num  73.5 75.5 66.5 64.5 72.4 ...
##  $ timbre_7_min            : num  -71.1 -65.8 -67.4 -63.7 -55.9 ...
##  $ timbre_7_max            : num  82.5 106.9 80.6 96.7 110.3 ...
##  $ timbre_8_min            : num  -52 -61.3 -59.8 -78.7 -56.5 ...
##  $ timbre_8_max            : num  39.1 35.4 46 41.1 37.6 ...
##  $ timbre_9_min            : num  -35.4 -81.9 -46.3 -49.2 -48.6 ...
##  $ timbre_9_max            : num  71.6 74.6 59.9 95.4 67.6 ...
##  $ timbre_10_min           : num  -126.4 -103.8 -108.3 -102.7 -52.8 ...
```

```
##  $ timbre_10_max           : num  18.7 121.9 33.3 46.4 22.9 ...
##  $ timbre_11_min           : num  -44.8 -38.9 -43.7 -59.4 -50.4 ...
##  $ timbre_11_max           : num  26 22.5 25.7 37.1 32.8 ...
##  $ Top10                   : int  0 0 0 0 0 0 0 0 0 1 ...
```

And below is the summary of contents for each variables.

```
summary(songs)
```

```
##       year          songtitle              artistname
##  Min.   :1990   Intro    :  15   Various artists: 162
##  1st Qu.:1997   Forever  :   8   Anal Cunt      :  49
##  Median :2002   Home     :   7   Various Artists:  44
##  Mean   :2001   Goodbye  :   6   Tori Amos      :  41
##  3rd Qu.:2006   Again    :   5   Eels           :  37
##  Max.   :2010   Beautiful:   5   Napalm Death   :  37
##                 (Other)  :7528   (Other)        :7204
##                 songID                   artistID    timesignature
##  SOALSZJ1370F1A7C75:   2   ARAGWS81187FB3F768: 222   Min.   :0.000
##  SOANPAC13936E0B640:   2   ARL14X91187FB4CF14:  49   1st Qu.:4.000
##  SOBDGMX12B0B80808E:   2   AR4KS8C1187FB4CF3D:  41   Median :4.000
##  SOBUDCZ12A58A80013:   2   AR0JZZ01187B9B2C99:  37   Mean   :3.894
##  SODFRLK13134387FB5:   2   ARZGTK71187B9AC7F5:  37   3rd Qu.:4.000
##  SOEJPOK12A6D4FAFE4:   2   AR95XYH1187FB53951:  31   Max.   :7.000
##  (Other)           :7562   (Other)           :7157
##  timesignature_confidence    loudness           tempo
##  Min.   :0.0000           Min.   :-42.451   Min.   :  0.00
##  1st Qu.:0.8193           1st Qu.:-10.847   1st Qu.: 88.86
##  Median :0.9790           Median : -7.649   Median :103.27
##  Mean   :0.8533           Mean   : -8.817   Mean   :107.35
##  3rd Qu.:1.0000           3rd Qu.: -5.640   3rd Qu.:124.80
##  Max.   :1.0000           Max.   :  1.305   Max.   :244.31
##
##  tempo_confidence      key         key_confidence       energy
##  Min.   :0.0000   Min.   : 0.000   Min.   :0.0000   Min.   :0.00002
##  1st Qu.:0.3720   1st Qu.: 2.000   1st Qu.:0.2040   1st Qu.:0.50014
##  Median :0.7015   Median : 6.000   Median :0.4515   Median :0.71816
##  Mean   :0.6229   Mean   : 5.385   Mean   :0.4338   Mean   :0.67547
##  3rd Qu.:0.8920   3rd Qu.: 9.000   3rd Qu.:0.6460   3rd Qu.:0.88740
##  Max.   :1.0000   Max.   :11.000   Max.   :1.0000   Max.   :0.99849
##
##      pitch          timbre_0_min     timbre_0_max    timbre_1_min
##  Min.   :0.00000   Min.   : 0.000   Min.   :12.58   Min.   :-333.72
##  1st Qu.:0.00300   1st Qu.: 0.000   1st Qu.:53.12   1st Qu.:-160.12
##  Median :0.00700   Median : 0.027   Median :55.53   Median :-107.75
##  Mean   :0.01082   Mean   : 4.123   Mean   :54.46   Mean   :-110.79
##  3rd Qu.:0.01400   3rd Qu.: 2.772   3rd Qu.:57.08   3rd Qu.: -59.71
##  Max.   :0.54100   Max.   :48.353   Max.   :64.01   Max.   : 123.73
##
##   timbre_1_max     timbre_2_min      timbre_2_max      timbre_3_min
##  Min.   :-74.37   Min.   :-324.86   Min.   : -0.832   Min.   :-495.36
##  1st Qu.:171.13   1st Qu.:-167.64   1st Qu.:100.519   1st Qu.:-226.87
##  Median :194.40   Median :-136.60   Median :129.908   Median :-170.61
```

```
##  Mean   :212.34   Mean   :-136.89   Mean   :136.673   Mean   :-186.11
##  3rd Qu.:239.24   3rd Qu.:-106.51   3rd Qu.:166.121   3rd Qu.:-131.56
##  Max.   :549.97   Max.   :  34.57   Max.   :397.095   Max.   : -21.55
##
##   timbre_3_max     timbre_4_min      timbre_4_max      timbre_5_min
##  Min.   : 12.85   Min.   :-207.07   Min.   : -0.651   Min.   :-262.48
##  1st Qu.:127.14   1st Qu.: -77.69   1st Qu.: 83.966   1st Qu.:-113.58
##  Median :189.50   Median : -63.83   Median :107.422   Median : -95.47
##  Mean   :211.81   Mean   : -65.28   Mean   :108.227   Mean   :-104.00
##  3rd Qu.:290.72   3rd Qu.: -51.34   3rd Qu.:130.286   3rd Qu.: -81.02
##  Max.   :499.62   Max.   :  51.43   Max.   :257.801   Max.   : -42.17
##
##   timbre_5_max     timbre_6_min       timbre_6_max     timbre_7_min
##  Min.   :-22.41   Min.   :-152.170   Min.   : 12.70   Min.   :-214.791
##  1st Qu.: 84.64   1st Qu.: -94.792   1st Qu.: 59.04   1st Qu.:-101.171
##  Median :119.90   Median : -80.418   Median : 70.47   Median : -81.797
##  Mean   :127.04   Mean   : -80.944   Mean   : 72.17   Mean   : -84.313
##  3rd Qu.:162.34   3rd Qu.: -66.521   3rd Qu.: 83.19   3rd Qu.: -64.301
##  Max.   :350.94   Max.   :   4.503   Max.   :208.39   Max.   :   5.153
##
##   timbre_7_max     timbre_8_min       timbre_8_max     timbre_9_min
##  Min.   : 15.70   Min.   :-158.756   Min.   :-25.95   Min.   :-149.51
##  1st Qu.: 76.50   1st Qu.: -73.051   1st Qu.: 40.58   1st Qu.: -70.28
##  Median : 94.63   Median : -62.661   Median : 49.22   Median : -58.65
##  Mean   : 95.65   Mean   : -63.704   Mean   : 50.06   Mean   : -59.52
##  3rd Qu.:112.71   3rd Qu.: -52.983   3rd Qu.: 58.46   3rd Qu.: -47.70
##  Max.   :214.82   Max.   :  -2.382   Max.   :144.99   Max.   :   1.14
##
##   timbre_9_max     timbre_10_min     timbre_10_max     timbre_11_min
##  Min.   :  8.415   Min.   :-208.82   Min.   : -6.359   Min.   :-145.599
##  1st Qu.: 53.037   1st Qu.:-105.13   1st Qu.: 39.196   1st Qu.: -58.058
##  Median : 65.935   Median : -83.07   Median : 50.895   Median : -50.892
##  Mean   : 68.028   Mean   : -87.34   Mean   : 55.521   Mean   : -50.868
##  3rd Qu.: 81.267   3rd Qu.: -64.52   3rd Qu.: 66.593   3rd Qu.: -43.292
##  Max.   :161.518   Max.   : -10.64   Max.   :192.417   Max.   :  -6.497
##
##  timbre_11_max        Top10
##  Min.   :  7.20   Min.   :0.0000
##  1st Qu.: 38.98   1st Qu.:0.0000
##  Median : 46.44   Median :0.0000
##  Mean   : 47.49   Mean   :0.1477
##  3rd Qu.: 55.03   3rd Qu.:0.0000
##  Max.   :110.27   Max.   :1.0000
##
```

In other markdown, we will peform an analysis and prediction modeling on the above dataset.