A Preliminary Analyzation of Grammy Album of the Year History with Spotify API

Sabrina Lem

```
## Warning: package 'knitr' was built under R version 4.0.1
```

Background and Data Extraction:
Grammy nominations and winners can be somewhat predictable. It's easy to place bets on who's going to get nominated or win based on streaming history, top-charts, or artist popularity. However, it is interesting to consider if the Recording Academy has a specific taste for music. This project intends to explore the trends in audio features for the nominations and winners of Album of the Year. To do this, I have generated a brief python application that connects to Spotify database. Spotify's open database returns various information about release history, artist popularity, streaming, history, etc. Amongst the features available from the Spotify Developer's page, you can request a track's audio features. Audio features are 13 Spotify-deemed and/or factual values that rate a track based on sonic characteristics. Each data point is an audio feature summary of each track in the album; Note Spotify does not offer this data- my Python application summarized each album (see a detailed explanation in the .py file). In addition to the audio features from Spotify I've included Spotify's Album name and ID, Spotify's Artist name and ID, the year of nomination, the number of tracks in the album, and a binary variable for if the album won in the year of respective nomination. The data frame includes every nomination from 1961 to 2021. Using this data frame I will look at how audio features have changed overtime and how those audio features can be used to model the Recording Academy's 'taste' for music.

Series in the data frame:
album_name: Album Name
album_id: Spotify's unique Album ID
artist_name: Primary artist of the Album
artist_id: Spotify's unique Artist ID
year: year of Grammy nomination number_track: Number of Tracks in the album
win: 1{won the year of nomination}, 0{otherwise}
danceability: Average measure from 0.0 to 1.0 of how danceable the tracks are (1.0 being most energetic)
energy: Average measure from 0.0 to 1.0 of how energetic the album's tracks are (1.00 being most energetic)
key: average of factor variable from -1.0 to 11.0 for the key of the album's tracks
loudness: Average overall loudness in decibels of the album's tracks.
mode: 1{major}, 0{minor}
speechiness: Average measure from 0.0 to 1.0 of whether the album's tracks are a speech recording (1.00 being exclusively speech)
acousticnes: Average measure from 0.0 to 1.0 of whether the album's track is acoustic (1.00 being most acoustic)
instrumentalnes: Average measure from 0.0 to 1.0 of whether the album's tracks are instrumental (1.00 being exclusively instrumental with no vocals)
liveness: Average from 0.0 to 1.0 of whether the album's tracks is a live recording/presence of an audience in recording (1.00 being most acoustic)
valence: Average measure from 0.0 to 1.0 of whether the album's tracks are ''positive and happy' sounding (1.00 being most valent sounding)
tempo:The average beats per minute of the the album's tracks
duration_ms:The average duration the album's track
time_signature: average of factor variable from 3 to 7 that indicates the album's tracks time signature

Data frame snippet:

```
library(readr)
master_data <- read_csv("master_data.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   .default = col_double(),
##   album_name = col_character(),
##   album_id = col_character(),
##   artist_name = col_character(),
##   artist_id = col_character()
## )

## See spec(...) for full column specifications.
```

```
df<-master_data[-c(1)]
df
```

```
## # A tibble: 308 x 20
##    album_name album_id artist_name artist_id  year number_tracks   win
##    <chr>      <chr>    <chr>       <chr>     <dbl>         <dbl> <dbl>
##  1 The Butto~ 7hX0R0e~ Bob Newhart 49mU7Sl9~  1961             6     1
##  2 Belafonte~ 26UI9qe~ Harry Bela~ 6Tw1ktF4~  1961            19     0
##  3 Nice 'n' ~ 2Xp6c8O~ Frank Sina~ 1Mxqyy3p~  1961            16     0
##  4 Puccini: ~ 3eTeuie~ Giacomo Pu~ 0OzxPXyo~  1961            37     0
##  5 Wild Is L~ 5k1MBVD~ Nat King C~ 7v4imS0m~  1961            14     0
##  6 Judy At C~ 3M0RdVA~ Judy Garla~ 0hItVPjw~  1962            26     1
##  7 Breakfast~ 53mCG3m~ Henry Manc~ 2EExdpjU~  1962            15     0
##  8 Genius + ~ 1VTSotS~ Ray Charles 1eYhYunl~  1962            10     0
##  9 Great Ban~ 52a7wxV~ The Si Zen~ 1eUYmq0o~  1962            12     0
## 10 The Nat K~ 3NoP1if~ Nat King C~ 7v4imS0m~  1962            36     0
## # ... with 298 more rows, and 13 more variables: danceability <dbl>,
## #   energy <dbl>, key <dbl>, loudness <dbl>, mode <dbl>, speechiness <dbl>,
## #   acousticness <dbl>, instrumentalness <dbl>, liveness <dbl>, valence <dbl>,
## #   tempo <dbl>, duration_ms <dbl>, time_signature <dbl>
```

What makes a nominee (summary of all nominations including winners):

```
df_nom <- df[-c(1:5,7)]
summary(df_nom)
```

```
##  number_tracks   danceability       energy            key           loudness
##  Min.   : 1.0   Min.   :0.230   Min.   :0.108   Min.   :2.00   Min.   :-21.37
##  1st Qu.:11.0   1st Qu.:0.477   1st Qu.:0.399   1st Qu.:4.41   1st Qu.:-13.44
##  Median :12.0   Median :0.561   Median :0.534   Median :5.08   Median :-10.17
##  Mean   :14.2   Mean   :0.556   Mean   :0.517   Mean   :5.15   Mean   :-10.56
##  3rd Qu.:15.2   3rd Qu.:0.642   3rd Qu.:0.638   3rd Qu.:5.77   3rd Qu.: -7.54
```

```
## Max.    :92.0   Max.   :0.844   Max.   :0.917   Max.   :9.00   Max.   : -2.50
##       mode        speechiness      acousticness    instrumentalness
## Min.   :0.143   Min.   :0.0286   Min.   :0.0001   Min.   :0.0000
## 1st Qu.:0.583   1st Qu.:0.0385   1st Qu.:0.1928   1st Qu.:0.0006
## Median :0.714   Median :0.0466   Median :0.3284   Median :0.0120
## Mean   :0.704   Mean   :0.0813   Mean   :0.3857   Mean   :0.0675
## 3rd Qu.:0.833   3rd Qu.:0.0755   3rd Qu.:0.5363   3rd Qu.:0.0726
## Max.   :1.000   Max.   :0.9438   Max.   :0.9565   Max.   :0.7257
##    liveness          valence           tempo        duration_ms
## Min.   :0.0627   Min.   :0.125   Min.   : 72.6   Min.   : 93034
## 1st Qu.:0.1342   1st Qu.:0.388   1st Qu.:108.5   1st Qu.:205304
## Median :0.1703   Median :0.498   Median :117.0   Median :239282
## Mean   :0.1988   Mean   :0.497   Mean   :116.3   Mean   :239447
## 3rd Qu.:0.2185   3rd Qu.:0.599   3rd Qu.:123.1   3rd Qu.:273878
## Max.   :0.9123   Max.   :0.917   Max.   :149.8   Max.   :427333
## time_signature
## Min.   :3.27
## 1st Qu.:3.81
## Median :3.91
## Mean   :3.87
## 3rd Qu.:4.00
## Max.   :4.10
```

What makes a winner (summary of all winners):

```
df_win<-df[df$win==1,]
df_win<-df_win[-c(1:5,7)]
summary(df_win)
```

```
## number_tracks   danceability       energy           key          loudness
## Min.   : 6.0   Min.   :0.230   Min.   :0.150   Min.   :2.64   Min.   :-18.68
## 1st Qu.:11.0   1st Qu.:0.483   1st Qu.:0.388   1st Qu.:4.57   1st Qu.:-13.92
## Median :13.0   Median :0.568   Median :0.514   Median :5.32   Median :-11.33
## Mean   :13.9   Mean   :0.551   Mean   :0.493   Mean   :5.32   Mean   :-11.26
## 3rd Qu.:15.0   3rd Qu.:0.617   3rd Qu.:0.618   3rd Qu.:6.24   3rd Qu.: -9.02
## Max.   :40.0   Max.   :0.808   Max.   :0.792   Max.   :8.44   Max.   : -4.22
##       mode        speechiness      acousticness    instrumentalness
## Min.   :0.154   Min.   :0.0303   Min.   :0.0416   Min.   :0.0000
## 1st Qu.:0.518   1st Qu.:0.0391   1st Qu.:0.2405   1st Qu.:0.0008
## Median :0.727   Median :0.0455   Median :0.4022   Median :0.0104
## Mean   :0.688   Mean   :0.0945   Mean   :0.4357   Mean   :0.0551
## 3rd Qu.:0.822   3rd Qu.:0.0672   3rd Qu.:0.6430   3rd Qu.:0.0729
## Max.   :1.000   Max.   :0.9438   Max.   :0.9292   Max.   :0.3736
##    liveness          valence           tempo        duration_ms
## Min.   :0.0717   Min.   :0.152   Min.   : 88.4   Min.   :127590
## 1st Qu.:0.1248   1st Qu.:0.380   1st Qu.:105.2   1st Qu.:215952
## Median :0.1584   Median :0.472   Median :114.6   Median :254571
## Mean   :0.2227   Mean   :0.484   Mean   :114.3   Mean   :251343
## 3rd Qu.:0.2110   3rd Qu.:0.578   3rd Qu.:121.1   3rd Qu.:282508
## Max.   :0.9123   Max.   :0.789   Max.   :139.7   Max.   :413925
## time_signature
## Min.   :3.33
## 1st Qu.:3.74
```

```
## Median :3.90
## Mean   :3.85
## 3rd Qu.:4.00
## Max.   :4.00
```

Audio Features Overtime:

RED: Winners
GREEN: Nominations (not including winners)

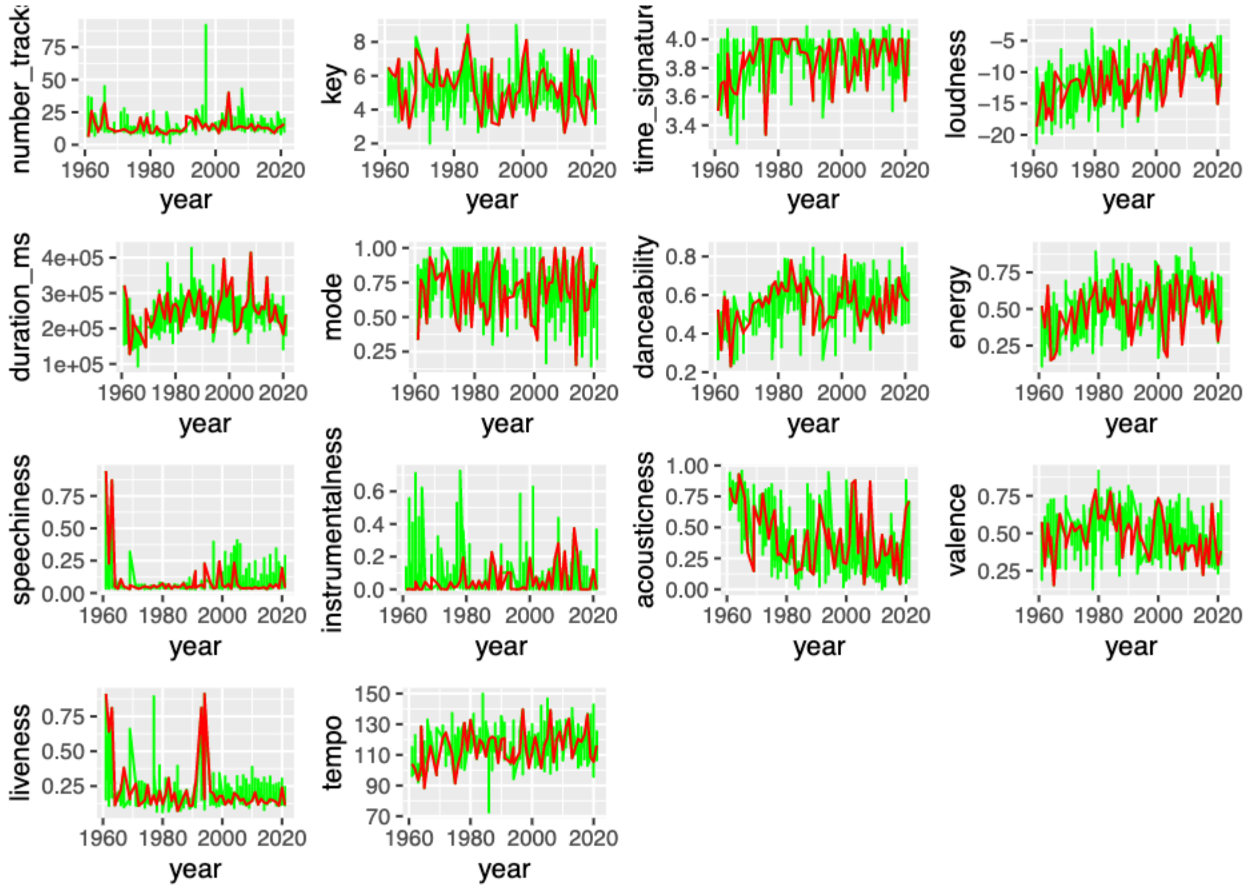
```r
df_nom1<-df[df$win==0,]
df_nom1 <-df[-c(1:4,7)]
df_win1<-df[df$win==1,]
df_win1<-df_win1[-c(1:4,7)]
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.1
```

```r
pl_count<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=number_tracks), color='green') +
  geom_line(data=df_win1, aes(x=year, y=number_tracks), color='red')
pl_key<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=key), color='green') +
  geom_line(data=df_win1, aes(x=year, y=key), color='red')
pl_ts<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=time_signature), color='green') +
  geom_line(data=df_win1, aes(x=year, y=time_signature), color='red')
pl_loudness<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=loudness), color='green') +
  geom_line(data=df_win1, aes(x=year, y=loudness), color='red')
pl_tempo<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=tempo), color='green') +
  geom_line(data=df_win1, aes(x=year, y=tempo), color='red')

pl_duration<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=duration_ms), color='green') +
  geom_line(data=df_win1, aes(x=year, y=duration_ms), color='red')
pl_mode<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=mode), color='green') +
  geom_line(data=df_win1, aes(x=year, y=mode), color='red')
pl_dance<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=danceability), color='green') +
  geom_line(data=df_win1, aes(x=year, y=danceability), color='red')
pl_energy<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=energy), color='green') +
  geom_line(data=df_win1, aes(x=year, y=energy), color='red')
pl_speech<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=speechiness), color='green') +
  geom_line(data=df_win1, aes(x=year, y=speechiness), color='red')
pl_instr<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=instrumentalness), color='green') +
  geom_line(data=df_win1, aes(x=year, y=instrumentalness), color='red')
pl_acoust<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=acousticness), color='green') +
  geom_line(data=df_win1, aes(x=year, y=acousticness), color='red')
pl_valence<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=valence), color='green') +
  geom_line(data=df_win1, aes(x=year, y=valence), color='red')
pl_liveness<-ggplot()+geom_line(data=df_nom1, aes(x=year, y=liveness), color='green') +
  geom_line(data=df_win1, aes(x=year, y=liveness), color='red')


library(gridExtra)
grid.arrange(pl_count, pl_key, pl_ts, pl_loudness, pl_duration,
             pl_mode, pl_dance,pl_energy,pl_speech,pl_instr,
             pl_acoust,pl_valence, pl_liveness, pl_tempo)
```



There are no outstanding trends in terms of the difference between those that win and those that were nominated. However, we can see here that there are some noticeable relationships between year and audio features. There is an upward trend in loudness, danceability, and energy. There is a downward trend in acousticness and liveness. This is an interesting facet to explore the social implications surrounding the decades. Other contributing factors to these trends could also be the technology surrounding the time (e.g. better technology implies louder decibel recordings).

OLS Regression:

```
mod <- lm(formula=win~loudness+duration_ms+danceability+energy+speechiness+
            instrumentalness+acousticness+valence+liveness+tempo+factor(year), data=df)
summary(mod)
```

```
##
## Call:
## lm(formula = win ~ loudness + duration_ms + danceability + energy +
##     speechiness + instrumentalness + acousticness + valence +
##     liveness + tempo + factor(year), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4886 -0.2255 -0.1499 -0.0227  0.9896
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.77e-01   6.20e-01   -1.25    0.211
## loudness          -1.28e-02   1.96e-02   -0.66    0.512
## duration_ms        1.61e-06   6.84e-07    2.36    0.019 *
## danceability       1.73e-01   4.34e-01    0.40    0.691
## energy             2.62e-01   5.58e-01    0.47    0.639
## speechiness       -2.96e-02   3.97e-01   -0.07    0.941
## instrumentalness -2.88e-01   2.42e-01   -1.19    0.235
## acousticness       3.73e-01   2.16e-01    1.73    0.085 .
## valence           -1.51e-01   3.61e-01   -0.42    0.677
## liveness           3.78e-01   2.97e-01    1.27    0.204
## tempo             -1.36e-03   2.79e-03   -0.49    0.626
## factor(year)1962  1.19e-01   2.70e-01    0.44    0.659
## factor(year)1963  1.67e-01   2.95e-01    0.57    0.572
## factor(year)1964  4.28e-01   2.82e-01    1.52    0.131
## factor(year)1965  2.38e-01   2.72e-01    0.88    0.381
## factor(year)1966  3.27e-01   2.88e-01    1.14    0.256
## factor(year)1967  2.21e-01   2.78e-01    0.79    0.427
## factor(year)1969  3.18e-01   2.55e-01    1.24    0.215
## factor(year)1971  3.27e-01   3.05e-01    1.07    0.284
## factor(year)1972  2.52e-01   2.90e-01    0.87    0.385
## factor(year)1973  4.40e-02   3.03e-01    0.15    0.885
## factor(year)1974  1.66e-01   2.89e-01    0.57    0.566
## factor(year)1975  2.93e-01   3.09e-01    0.95    0.344
## factor(year)1976  2.41e-01   2.93e-01    0.82    0.412
## factor(year)1977  1.81e-01   3.04e-01    0.60    0.551
## factor(year)1978  2.77e-01   2.98e-01    0.93    0.352
## factor(year)1979  2.87e-01   2.95e-01    0.97    0.331
## factor(year)1980  2.69e-01   3.03e-01    0.89    0.375
## factor(year)1981  2.52e-01   2.92e-01    0.86    0.389
## factor(year)1982  2.15e-01   2.99e-01    0.72    0.473
## factor(year)1983  3.08e-01   3.02e-01    1.02    0.308
## factor(year)1984  3.34e-01   3.05e-01    1.10    0.274
## factor(year)1985  2.80e-01   3.05e-01    0.92    0.360
## factor(year)1986  1.03e-01   3.03e-01    0.34    0.734
## factor(year)1987  1.70e-01   2.99e-01    0.57    0.569
## factor(year)1988  2.16e-01   3.02e-01    0.71    0.476
## factor(year)1989  2.14e-01   3.01e-01    0.71    0.479
## factor(year)1990  3.15e-01   3.06e-01    1.03    0.303
## factor(year)1991  2.31e-01   2.64e-01    0.87    0.383
## factor(year)1993  1.48e-01   2.94e-01    0.50    0.615
## factor(year)1994  1.81e-01   2.57e-01    0.70    0.482
## factor(year)1996  2.46e-01   3.04e-01    0.81    0.420
## factor(year)1997  2.89e-01   3.00e-01    0.96    0.336
## factor(year)1998  2.05e-01   2.98e-01    0.69    0.492
## factor(year)1999  2.70e-01   3.01e-01    0.90    0.370
## factor(year)2000  2.83e-01   2.99e-01    0.95    0.345
## factor(year)2001  2.47e-01   3.00e-01    0.82    0.412
## factor(year)2002  2.97e-01   3.08e-01    0.96    0.336
## factor(year)2003  2.72e-01   2.98e-01    0.91    0.361
## factor(year)2004  3.14e-01   2.99e-01    1.05    0.295
## factor(year)2005  3.09e-01   2.97e-01    1.04    0.300
## factor(year)2006  3.81e-01   3.06e-01    1.25    0.214
## factor(year)2007  2.79e-01   3.09e-01    0.90    0.367
```

```
## factor(year)2008  2.42e-01   3.00e-01    0.81    0.421
## factor(year)2009  3.25e-01   3.08e-01    1.06    0.292
## factor(year)2010  3.15e-01   3.11e-01    1.01    0.312
## factor(year)2011  3.39e-01   3.13e-01    1.08    0.279
## factor(year)2012  3.36e-01   3.13e-01    1.07    0.285
## factor(year)2013  3.66e-01   2.99e-01    1.23    0.221
## factor(year)2014  2.51e-01   3.08e-01    0.82    0.415
## factor(year)2015  2.63e-01   3.03e-01    0.87    0.386
## factor(year)2016  2.72e-01   3.04e-01    0.90    0.372
## factor(year)2017  2.93e-01   3.04e-01    0.97    0.335
## factor(year)2018  3.39e-01   3.03e-01    1.12    0.265
## factor(year)2019  2.46e-01   2.91e-01    0.85    0.398
## factor(year)2020  2.11e-01   2.78e-01    0.76    0.447
## factor(year)2021  2.42e-01   2.73e-01    0.88    0.377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.43 on 241 degrees of freedom
## Multiple R-squared:  0.0785, Adjusted R-squared:  -0.174
## F-statistic: 0.311 on 66 and 241 DF,  p-value: 1
```

This is a log linear regression. Each coefficient estimates the probability increase/decrease in winning if there is a 1 unit change in audio feature rating. However,even with this fixed effect model, using a dummy variable for year, it seems that audio features are not significant predictors for if an album will win the Album of the Year category. This could indicate that the Recording Academy has a fairly diverse taste in music: It can be observed that the album nominations involves a selection of albums that are diverse in genre and thus in audio features. Or simply, the audio features supplied by Spotify are fairly irrelevant to capturing the complex aspects of a track.

Since there is not much statistical evidence from this preliminary analysis, it may be more interesting to apply this python application elsewhere. Perhaps the python application can be used to compare user listening habits rather than make predictions. The python application is built to be changed easily to accept playlist input instead of albums. So we can use playlists as a way to sample a user's listening habits. Questions could involve: To what extent is music taste similar to my peers? How has listening changed over the years for a user (using playlists as a timestamp)? Going forward I intend to apply and adapt my Python application to suit these queries.