Search

Home

Library

Regression Final Project:

Spotify vs. Apple Songs in Playlist!

Sashwath Chetlur, Matthew Rothman,
Kayla Ventura, Tiffany Xu

Table of Contents

Questions / Goals

1. **Popularity**
   What factors make a song appear more in number of Spotify playlists? Apple playlists?

2. **Spotify**
   Create one MLR model with the predictor being #_in_spotify_playlists

3. **Apple**
   Create one MLR model with the predictor being #_in_apple_playlists

Bias / Issues in Data

Genre

Disproportionate

Observations per genre are disproportionate

Streams

Missing Information

Total streams are only listed for Spotify

Playlist

Disproportionate

There are a disproportionate number of Spotify vs Apple users/playlists

Search

Home

Library

Dataset & Cleaning

Original Dataset

- Top Spotify Songs (kaggle.com)
- Contains list of popular songs for spotify and apple
- 953 Observations
- Original 24 columns

Data Cleaning

- Manually added a feature “genre” to each observation
- Removed all observations with null values and features that were not wanted
- Remaining observations = 804
- Remaining columns = 14

Regressors: Categorical

Genre

Categorical: What type of genre the song is

Key

Categorical: The note of the song

Mode

Categorical: Major or Minor

Regressors: Numerical

BPM
Numerical: Beats per minute

Valence
Numerical: The musical positivity of the song

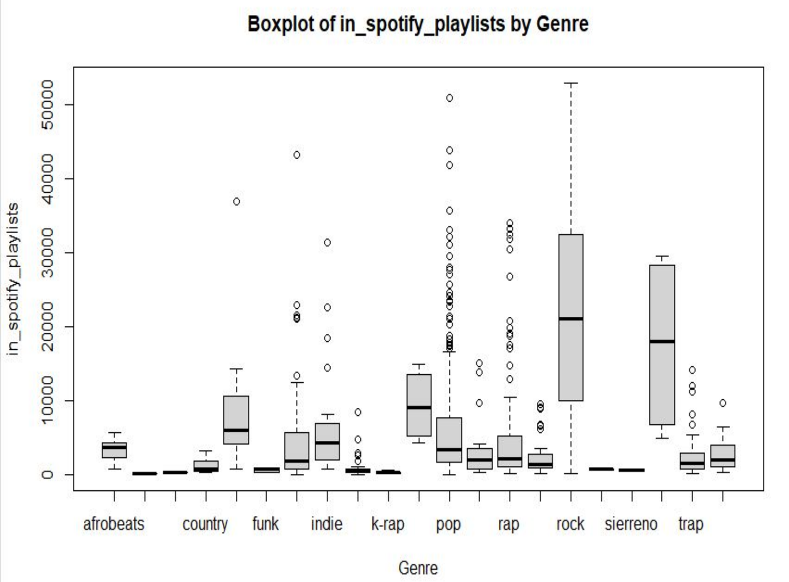
Danceability
Numerical: How danceable the song is

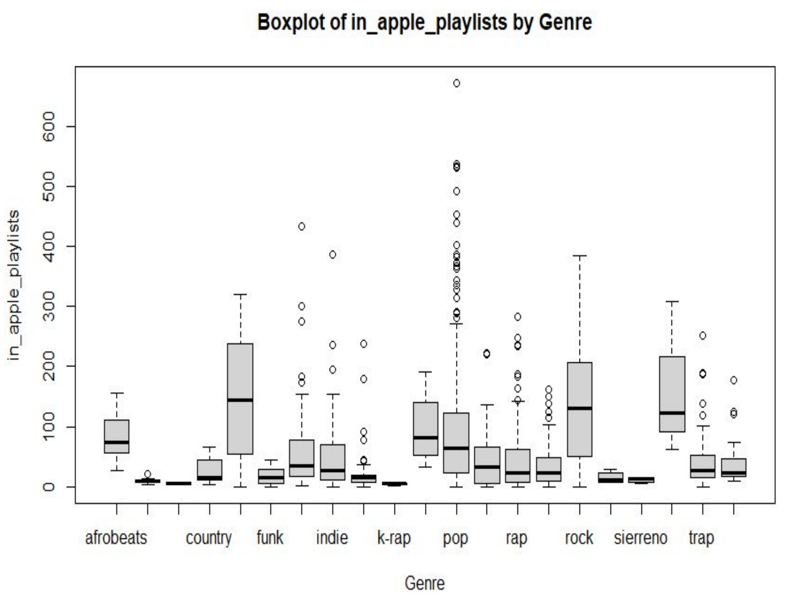
Energy
Numerical: How upbeat the song is

Acousticness
Numerical: Measure of how acoustic the song is

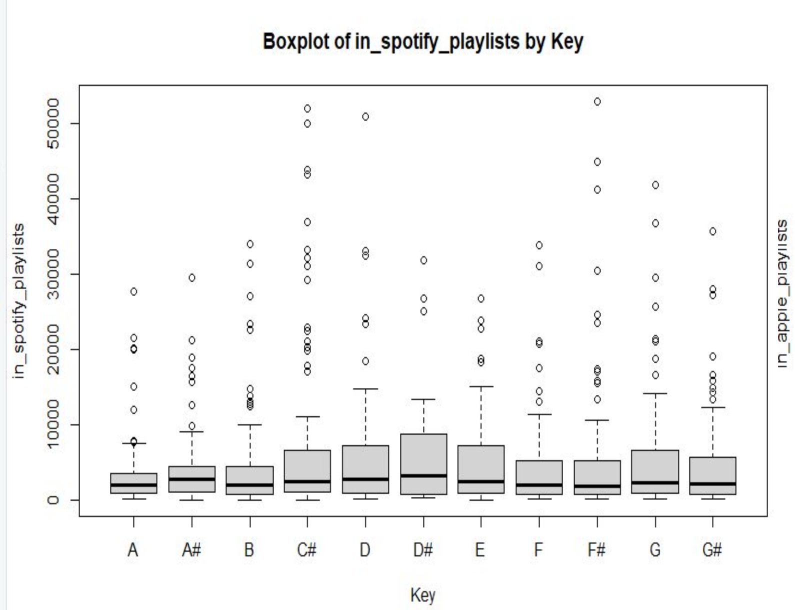
Instrumentalness
Numerical: Measure of how instrumental the song is

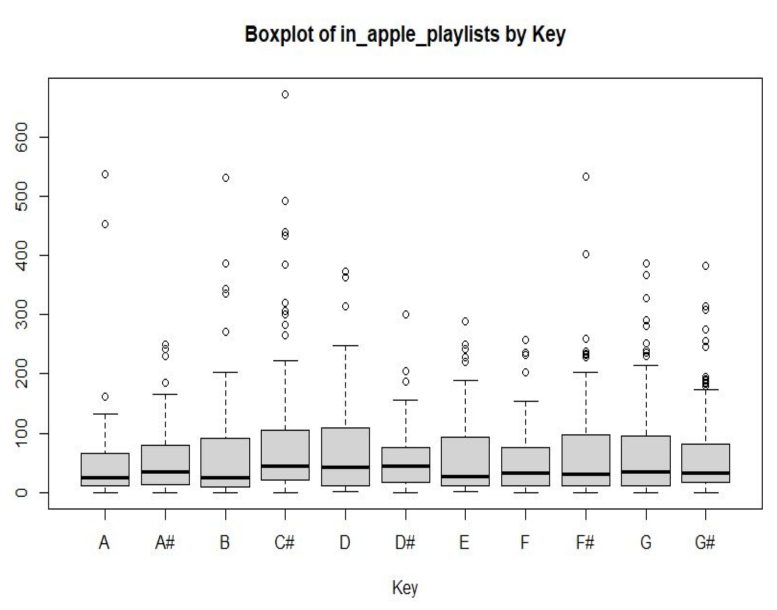
Spotify Vs Apple: Genre Boxplot



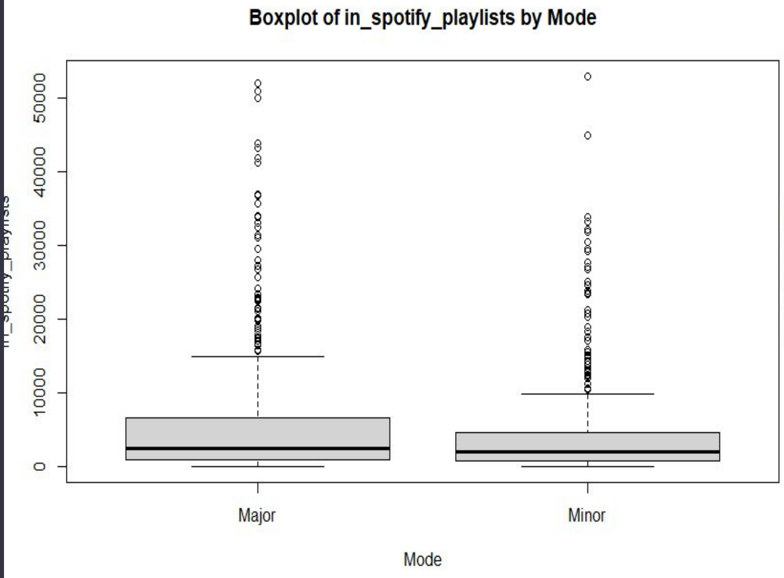


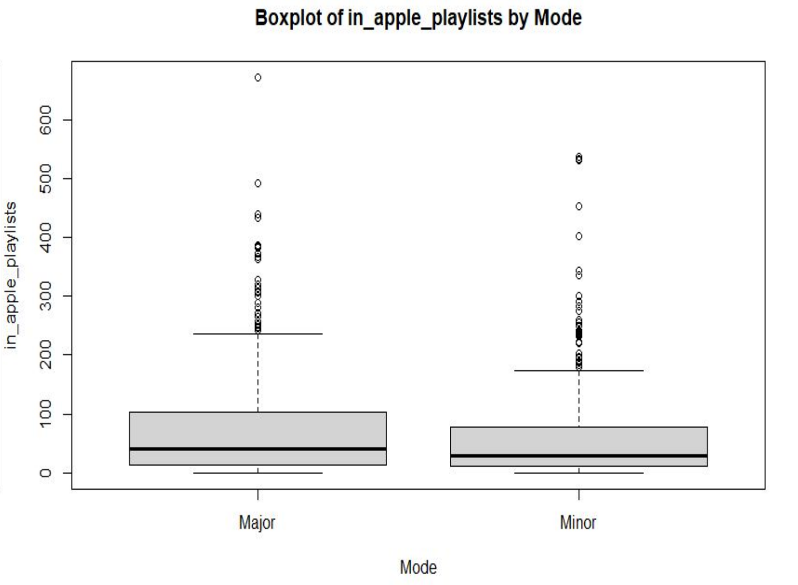
Spotify Vs Apple: Key Boxplot





Spotify vs Apple : Mode Boxplot





Boxplot Summary

- In both Apple and Spotify, the pop genre had the most outliers
- Between the two streaming services the concentration of genre observations varied greatly
- Comparing the observations of key between the streaming services shows a consistent distribution for C#, but other keys have a high variability in both Spotify and Apple
- Comparing the major and minor modes, major and minor in Spotify has similar concentration of points while Apple has a wider distribution for major
- We decided to keep all categorical variables because of the variability between each regressor

Spotify: Backward Model

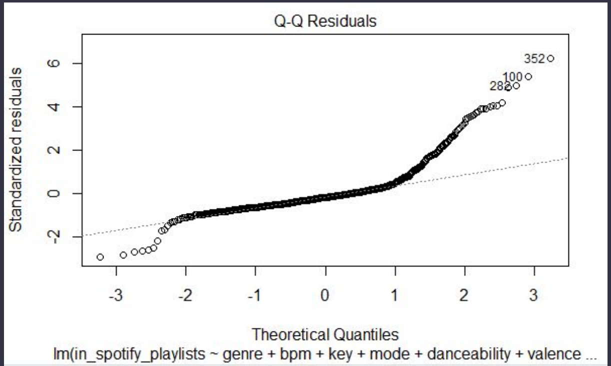
```
Call:
lm(formula = in_spotify_playlists ~ genre + energy, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-20938  -3603  -1416    788  43710

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          -20.00    2868.26  -0.007 0.994437
genreBollywood     -3356.41    3760.28  -0.893 0.372349
genreCorrido       -3144.54    4119.00  -0.763 0.445441
genrecountry       -2633.11    3348.34  -0.786 0.431875
genreedm            4722.99    3262.62   1.448 0.148128
genrefunk          -2229.73    3469.26  -0.643 0.520600
genrehip hop        1601.69    2800.52   0.572 0.567537
genreindie          3733.22    3011.17   1.240 0.215425
genreK-Pop         -2868.64    2845.88  -1.008 0.313767
genrek-rap         -3189.85    4119.10  -0.774 0.438925
genrelounge         7696.46    4440.37   1.733 0.083438 .
genrepop            3152.88    2691.58   1.171 0.241799
genreR&B             206.70    3053.46   0.068 0.946046
genrerap            2205.01    2744.22   0.804 0.421924
genrereggaeton     -1466.43    2798.28  -0.524 0.600394
genrerock          18896.60    2943.80   6.419 2.38e-10 ***
genresertanejo     -3309.37    4125.48  -0.802 0.422693
genresierreno      -2254.09    4411.32  -0.511 0.609509
genresoul          14742.01    4411.89   3.341 0.000873 ***
genretrap           -701.10    2817.30  -0.249 0.803538
genreUrban Latino   -781.87    3070.83  -0.255 0.799089
energy                51.85      16.63   3.118 0.001889 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

s: 7034 on 782 degrees of freedom
Multiple R-squared: 0.2648,
Adjusted R-squared: 0.2451
F-statistic: 13.41 on 21 and 782 DF,  p-value: < 2.2e-16
```

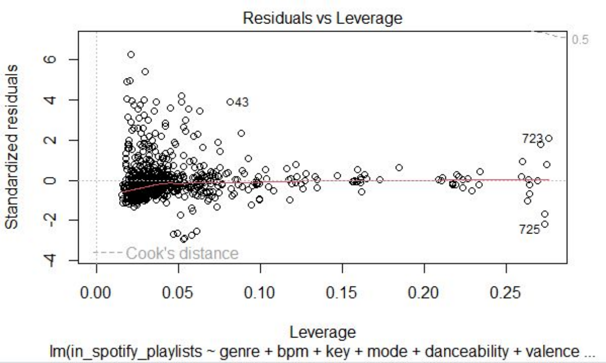
- Alpha: 0.05

Interpretation

- $Y = -20.00 \pm \text{genrecoeff}x_1 + 51.85x_2$
- Genre is the first regressor added, with energy added subsequently
- Energy has a low p-value (0.001889), which suggests that it is statistically significant
- The F-statistic (13.41) suggests that the model's predictors collectively have a significant effect on the dependent variable
- R^2 value is one of the higher model options

Spotify: Model Plots





- Light-tailed distribution
- We have more extreme values than would be expected of a normal distribution
- There are not any leverage points according to the Residuals Vs. Leverage plot.

Apple: Backward Model

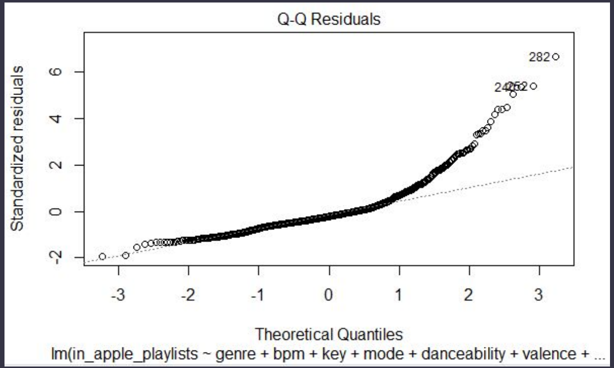
```
Call:
lm(formula = in_apple_playlists ~ genre + bpm + danceability +
    energy, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-147.44  -43.11  -18.91   15.28  560.91

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -12.1194    38.5895  -0.314  0.75356
genreBollywood    -77.5005    43.7986  -1.769  0.07721 .
genreCorrido      -88.8685    48.2114  -1.843  0.06566 .
genrecountry      -64.0704    39.2260  -1.633  0.10280
genreedm           46.9947    38.0716   1.234  0.21743
genrefunk         -73.8543    40.5913  -1.819  0.06922 .
genrehip hop      -26.4415    32.7012  -0.809  0.41900
genreindie        -14.3762    35.3226  -0.407  0.68412
genreK-Pop        -68.8622    33.2267  -2.072  0.03855 *
genrek-rap        -80.5834    48.0163  -1.678  0.09370 .
genrelounge        30.6312    51.8840   0.590  0.55511
genrepop           14.7070    31.4707   0.467  0.64040
genreR&B          -26.2833    35.7037  -0.736  0.46186
genrerap          -38.9141    32.0292  -1.215  0.22475
genrereggaeton    -56.9888    32.7049  -1.743  0.08181 .
genrerock          56.6562    34.4987   1.642  0.10094
genresertanejo    -81.9132    48.1904  -1.700  0.08957 .
genresierreno     -75.5975    51.5334  -1.467  0.14279
genresoul          72.9664    51.4257   1.419  0.15634
genretrap         -48.0112    32.8944  -1.460  0.14482
genreUrban Latino -51.2284    35.8521  -1.429  0.15344
bpm                 0.2329     0.1060   2.197  0.02834 *
danceability        0.5619     0.2234   2.515  0.01210 *
energy              0.5072     0.1952   2.598  0.00955 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

s: 81.92 on 780 degrees of freedom
Multiple R-squared: 0.1622,
Adjusted R-squared: 0.1375
F-statistic: 6.565 on 23 and 780 DF,  p-value: < 2.2e-16
```

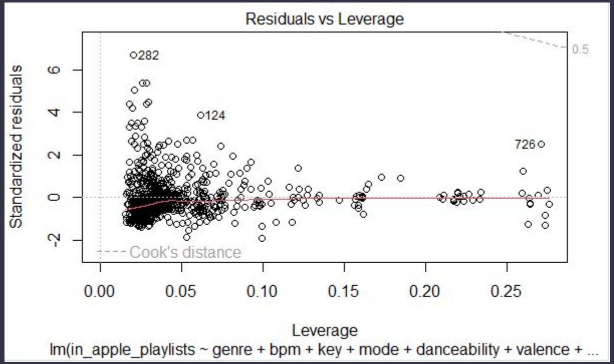
- Alpha = 0.05

Interpretation

- $Y = -12.1194 \pm \text{genrecoeff}x_1 + 0.2329x_2 + 0.5619x_3 + 0.5072x_4$
- Genre is the first regressor added, with BPM, danceability, and energy added subsequently
- The F-statistic (6.565) suggests that the model’s predictors collectively have a significant effect on the dependent variable
- Energy has the lowest p-value (0.00955), which suggests that it is statistically significant
- R^2 value is one of the higher model options

Apple: Model Plots





- Light Right Tail Distribution
- No leverage points –all points far from Cook's D
- Distribution shows more extreme values than a normal distribution

Comparing the Two Models

- Spotify only deemed two regressors as significant for the prediction of amount of playlists a song will appear in
- Apple picked four regressors that were significant to our predictor
- Both step functions deemed genre the most significant, adding that regressor first to the model
- While energy was added in both models bpm and danceability were deemed more significant in the model for Apple playlists
- Overall the Spotify model deemed the only necessary regressors to be energy and genre while the Apple model selected genre, bpm, danceability and energy

Final Interpretation

- H_0: All β_i are = 0
- H_1: At least one $\beta_i \neq 0$
- Both backward models for Spotify and Apple have a statistically significant F-statistic, which suggest that all of the predictors collectively have a significant effect on the dependent variable (in_spotify_playlists & in_apple_playlists)
- The p-values for both the Spotify and Apple backward models are 2.2e-16, which is very close to 0. This indicates an extremely high level of significance.
- **Final:** As both the p-values are 2.2e-16, which is less than the alpha of 0.05, we reject the null hypothesis, and conclude that models are both statistically significant.

Search

Home

Library

Thanks!

Do you have any questions?

Sashwath Chetlur, Matthew Rothman,
Kayla Ventura, Tiffany Xu

Search

Home

Library

Appendix

Additional Work

All additional work done that was not included in the final presentation

Search

Home

Library

Regression Final Project

- Unknown what the original dataset's focus was.
- Unknown how regressors like danceability were created
- Compared stepwise model selection with backward. For #_in_spotify_playlists they both ended up with the same features at a 10% and 5% level. However, when comparing with apple, the backward was more stringent at both 10% and 5%. We decided to stay with backward selection to get our final model
- There does not seem to be a transformation required for linearization
- Tried Logistic Regression model on data by converting a category into 1s and 0s
- Relationship produced a strong nonlinear relationship
- Tried a few different transformation to the model such as: sqrt and log
- None of the transformations produced a linear relationship(decided to move away from this approach)