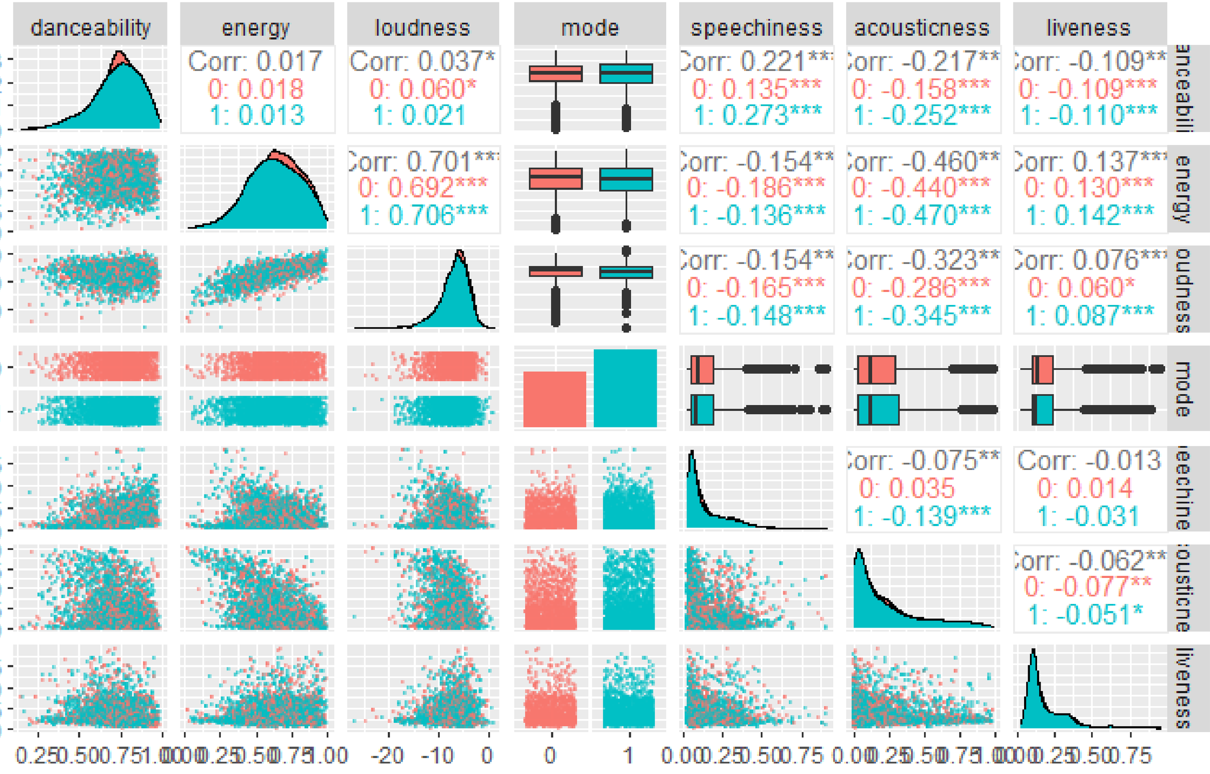
The TikTok Influencer's Guide

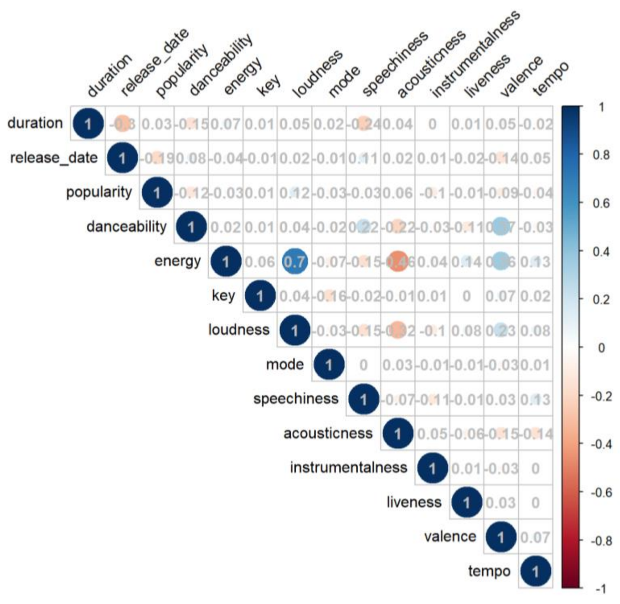
Yikai Cao, Haoran Guo, Yishu Ji, Department of Industrial Engineering, Tsinghua University

Introduction & EDA

Nowadays, TikTok is widespread used across the world, with the songs in it get more and more popular. We are interested in how the popularity of songs correlated with their other features.

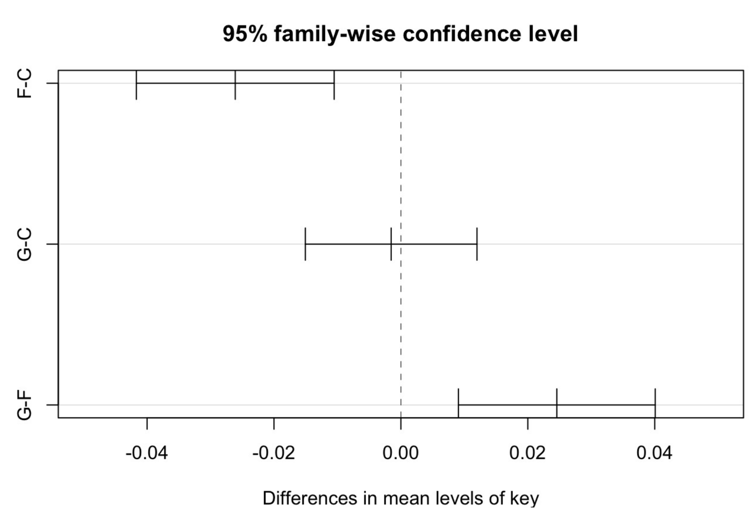


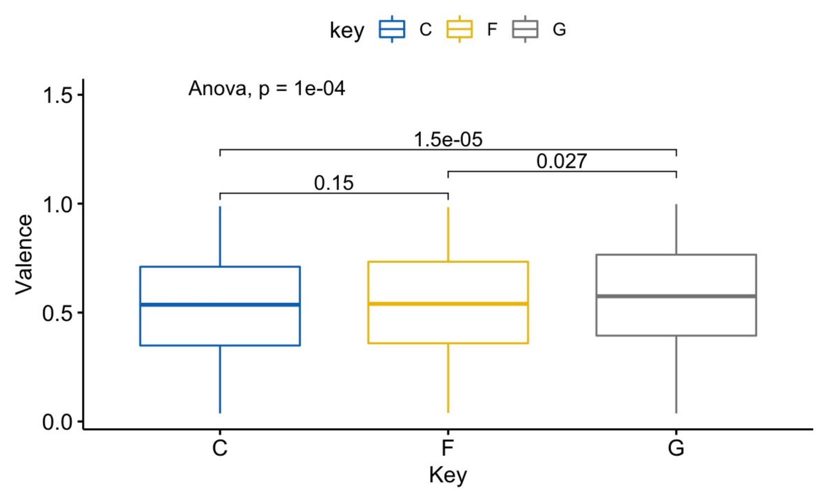
From the pairwise plot above, we can roughly see that:
(1) There are no significant differences between mode 0 and 1;
(2) Energy and loudness are **positively correlated**;
(3) Energy and Acousticness are **negatively correlated**;
(4) The distributions of speechiness, acousticness, and liveness are **right-skewed**.



The **correlation matrix** of all variables shows no obvious linear correlation between most of them. The most significant positive correlation lies in energy and loudness with a coefficient of 0.7, and the most significant negative correlation lies in energy and acousticness with a coefficient of -0.46.

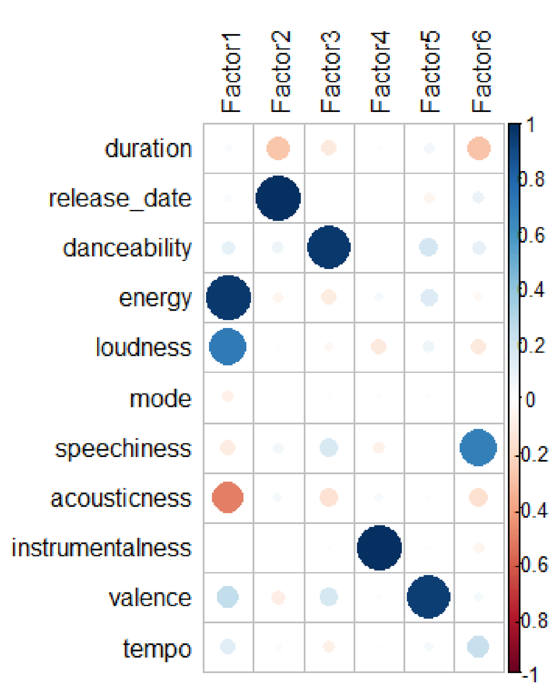
ANOVA on key

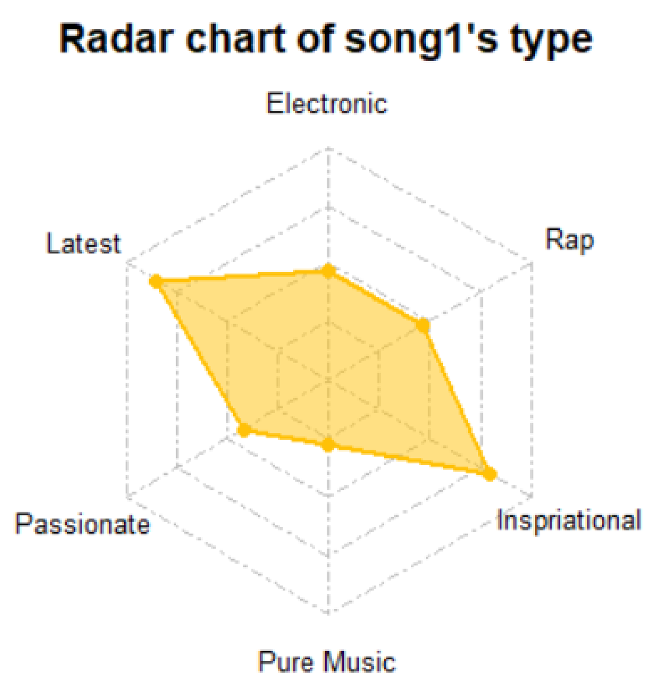




By applying ANOVA, the deviations in songs' features according to different key groups are analyzed. It can be summarized that the danceability and valence significantly vary from group to group, which is in consistence with basic musical knowledge

Analysis on style





We use **MLE method** to decompose the explanation variables into 6 **latent factors**, with the loadings given on the heatmap above. The six factors are named as 'Electronic', 'Rap', 'Inspirational', 'Pure Music', 'Passionate', 'Latest'. For each song, we can compute its **factor scores** and show the corresponding **radar chart**.

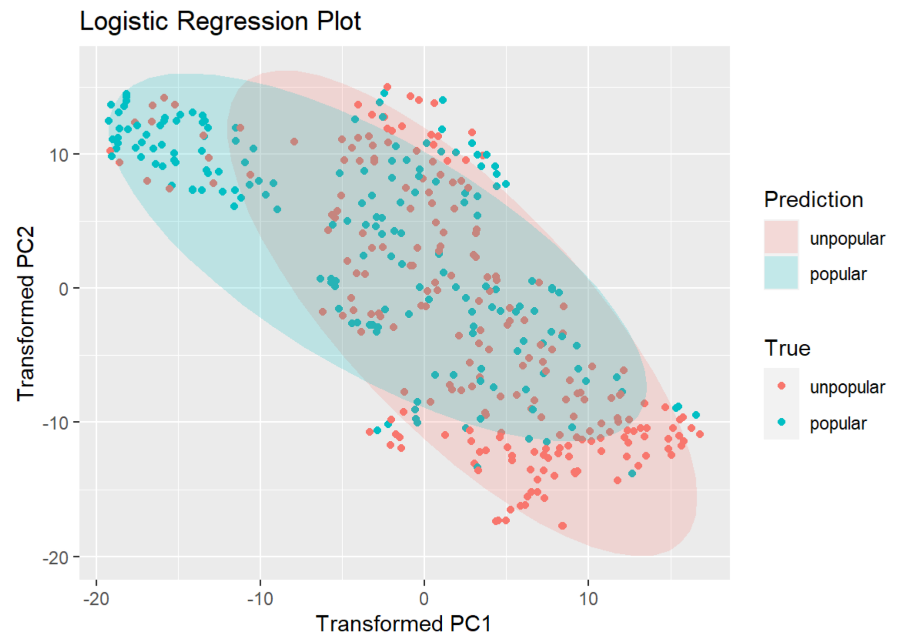
Prediction on Popularity

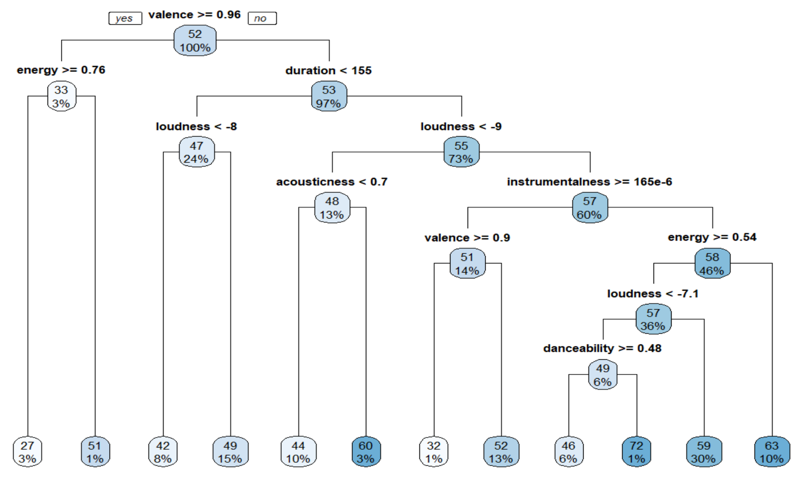
```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.522e+02  5.657e+00  26.898  < 2e-16 ***
duration          -1.763e-02  6.463e-03  -2.728  0.00639 **
release_date      -2.623e-03  1.874e-04 -14.000  < 2e-16 ***
danceability      -1.546e+01  3.007e+00  -5.140 2.87e-07 ***
energy            -2.317e+01  3.248e+00  -7.135 1.13e-12 ***
loudness           2.375e+00  1.747e-01  13.595  < 2e-16 ***
mode1             -2.239e+00  7.343e-01  -3.049  0.00231 **
speechiness        5.102e+00  3.080e+00   1.656  0.09775 .
acousticness       5.252e+00  1.815e+00   2.894  0.00383 **
instrumentalness -1.072e+01  2.455e+00  -4.368 1.29e-05 ***
valence           -9.065e+00  1.825e+00  -4.967 7.06e-07 ***
tempo             -2.541e-02  1.429e-02  -1.778  0.07546 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.53 on 4250 degrees of freedom
Multiple R-squared:  0.1082,    Adjusted R-squared:  0.1059
F-statistic: 46.89 on 11 and 4250 DF,  p-value: < 2.2e-16
```

Despite the significance of some variables, the direct **multiple linear regression** doesn't perform well since the best possible adjusted R^2 being 0.1082. Hence, we will conduct slices analysis later, i.e. try regression on the subset of data divided by playlist name or artist name.

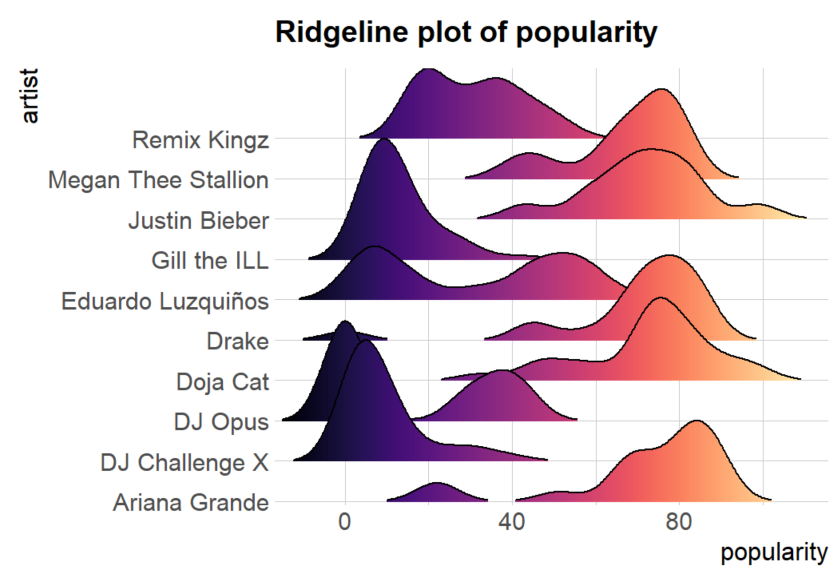
We conduct the **logistic regression** to predict whether a song will be popular (popular > 60) in TikTok. The plot based on the **transformed principal components** on the right side shows the training result. The color of points in the plot indicates the true value of 'popular or not', and the ellipses represent the prediction region.



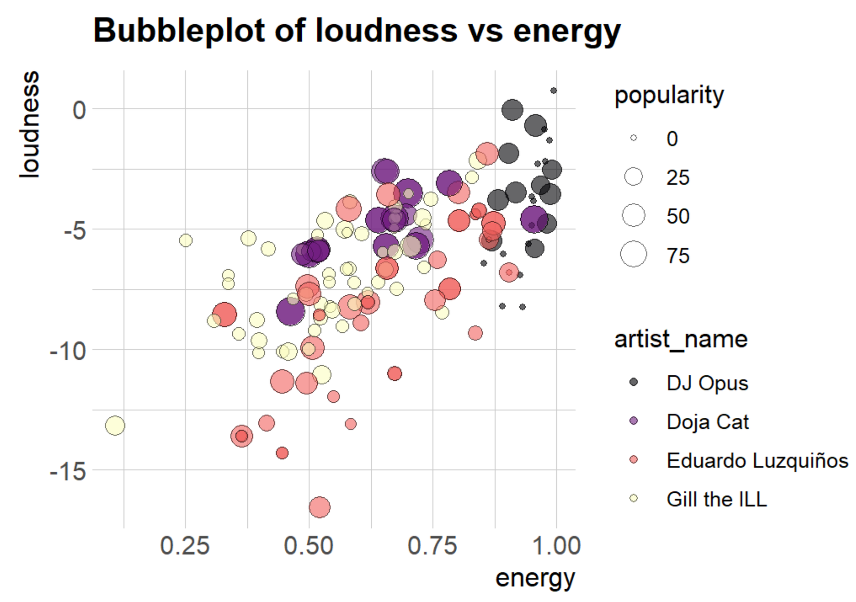


The **decision tree** divides variables into multiple intervals and gives the range of variable values to obtain more popularity.

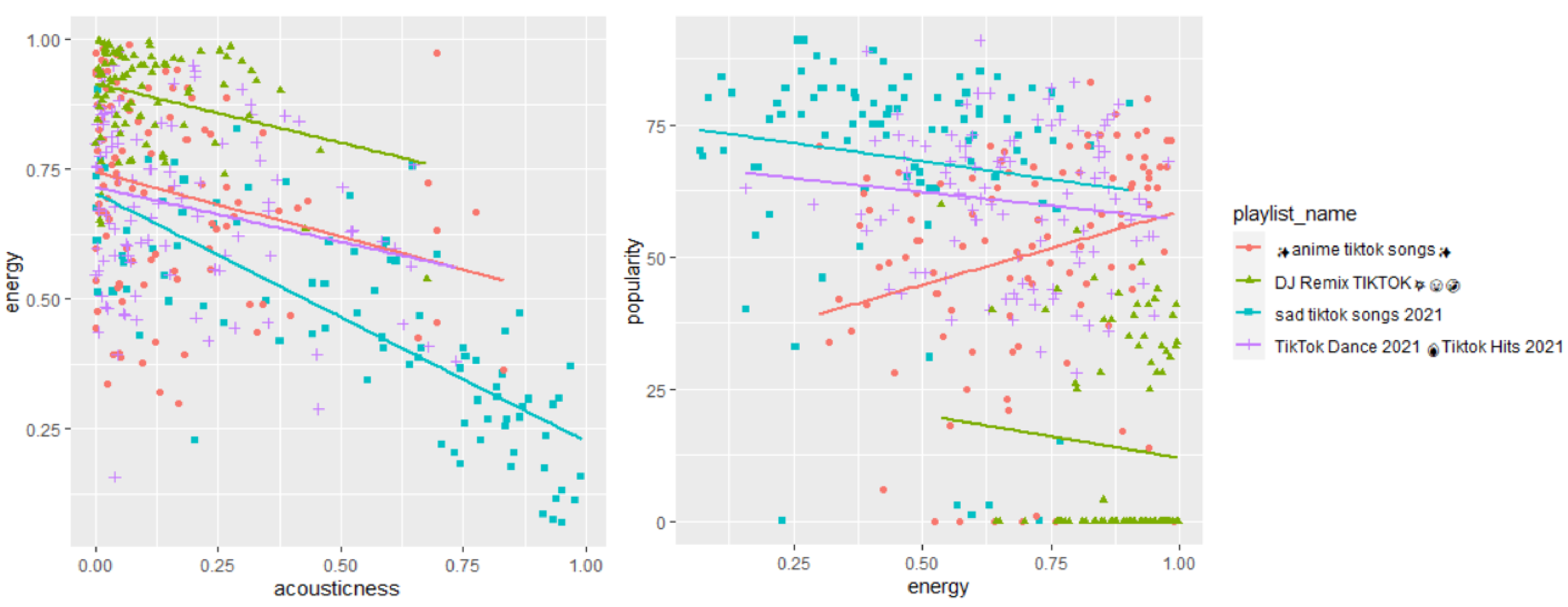
Analysis on category



The **ridgeline plot** of popularity indicates distinctions between artists. Grammy Award-winning artists enjoy higher popularity, while DJ artists gain lower popularity.



The **bubble plot** of loudness and energy shows an obvious positive correlation between the two variables. We could use this plot to suggest that the songs of DJ Opus possess higher energy and loudness. Besides, the songs of Doja Cat enjoy higher popularity, while Gill the ILL gains lower popularity.



We select some representative playlists and analyze some properties of them. Some of them show significance **interaction effects**. For instance, acousticness and energy are negatively correlated for sad songs only; popularity is slightly negatively correlated with energy for many songs, but that's not the case for anime songs.