TikTok Trending Tracks

In-process Inspection of Project

Peisen Li, Junhao Yu, Yi Zheng

Weiyang College, Tsinghua University

2022.5.14

1. Exploratory Data Analysis
2. Model selection
3. Some questions

Overview of data

- DANCE dimension: 1891×21
- OPM dimension: 259×21
- general dimension: 770×21
- PHILIPPINES dimension: 1725×21

Overview of predictors

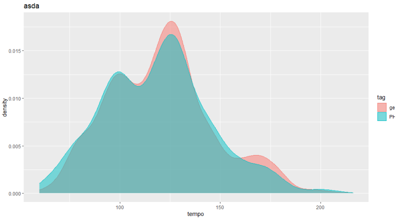
Some for identification:

- track_id, track_name, artist_id, etc.

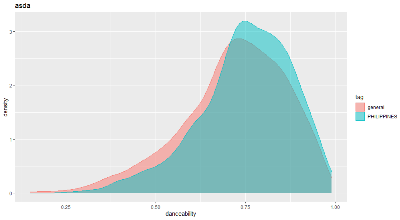
Some for analysis:

- danceability, energy, key, etc. (as independent variables)
- popularity (as dependent/response variable)

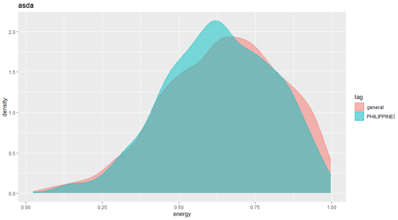
Bind "general" case with "PHILIPPINES"



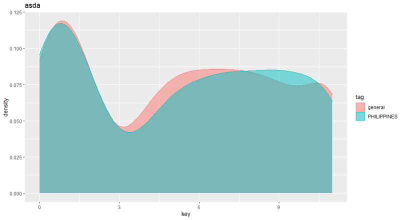
Tempo



Danceability

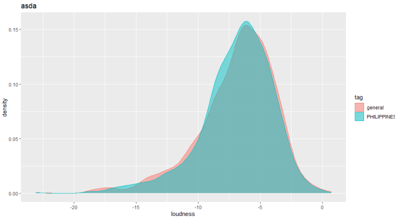


Energy

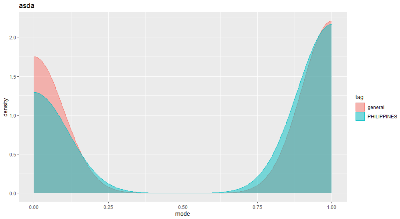


Key

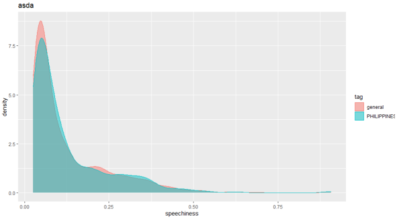
Bind "general" case with "PHILIPPINES"



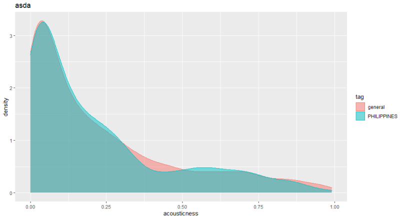
Loudness



Mode

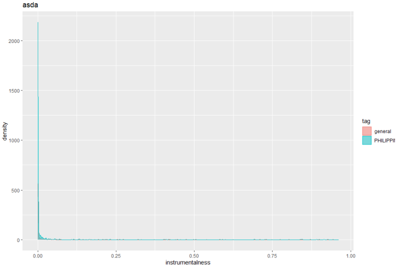


Speechiness

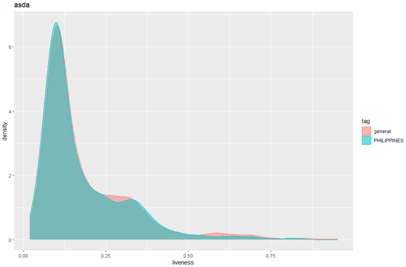


Acousticness

Bind "general" case with "PHILIPPINES"

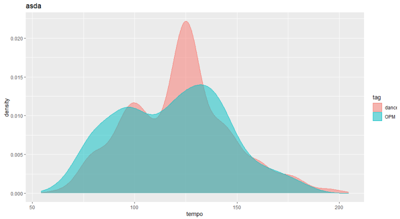


Instrumentalness

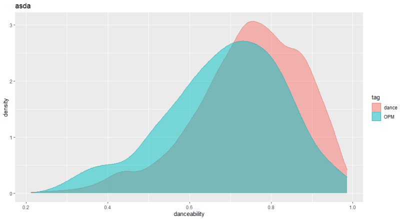


Liveness

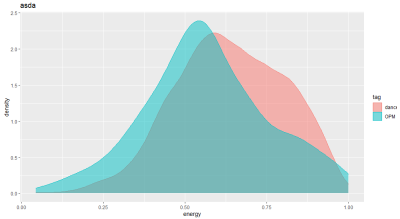
Comparison between "dance" and "OPM"



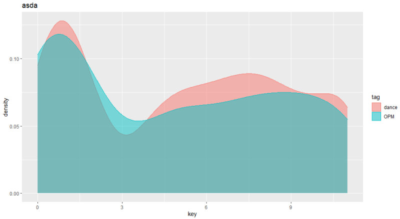
Tempo



Danceability

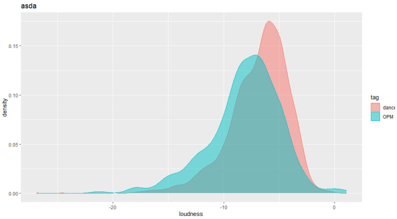


Energy

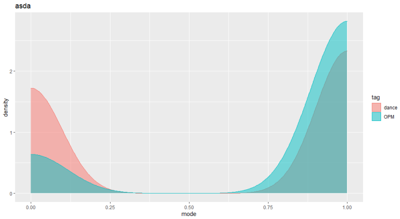


Key

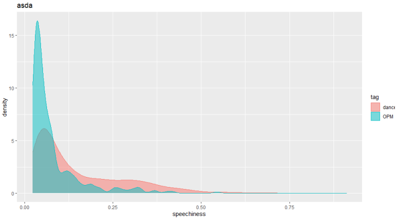
Comparison between "dance" and "OPM"



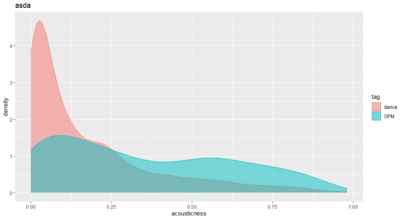
Loudness



Mode

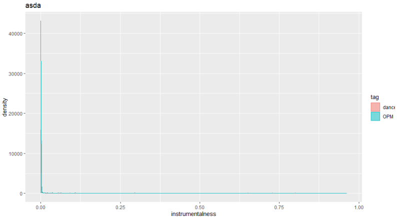


Speechiness

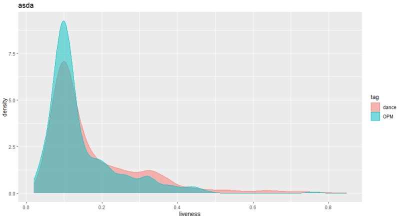


Acousticness

Comparison between "dance" and "OPM"

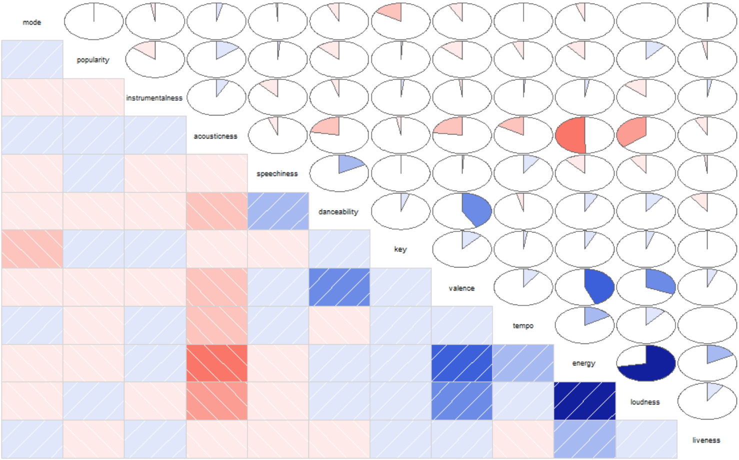


Instrumentalness

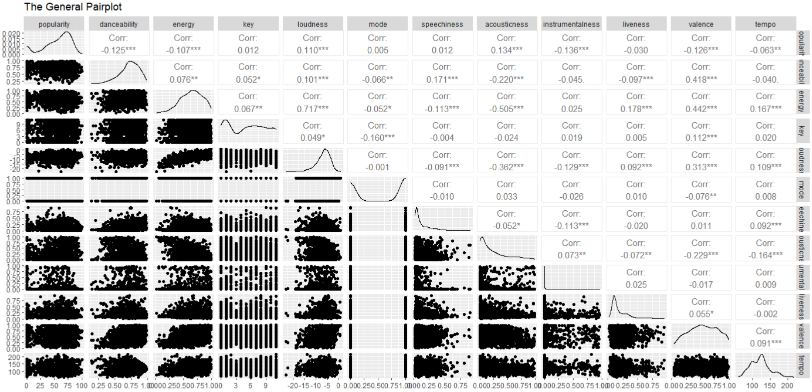


Liveness

Correlation plot of general



Pairplot of general



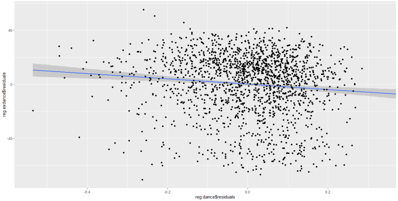
Main goal

Our main goal is to fit a model to predict the 'popularity' using the other predictors.

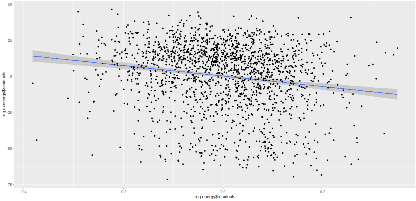
1. Exploratory Data Analysis
2. Model selection
   - Linear model
   - SVM regression
   - Tree method
3. Some questions

- We have tried ordinary linear model.
- To explore whether those predictors are really linearly dependent, we drew partial residuals plots.

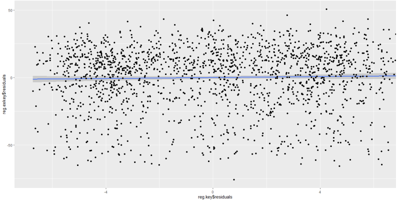
Linear Regression



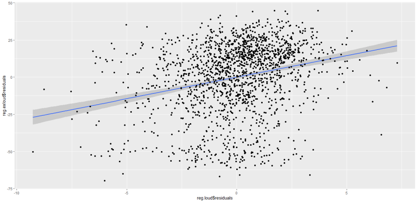
Danceability



Energy

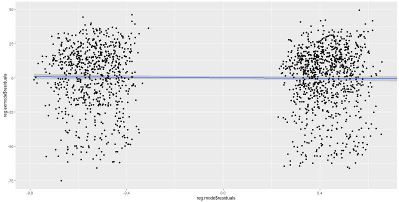


Key

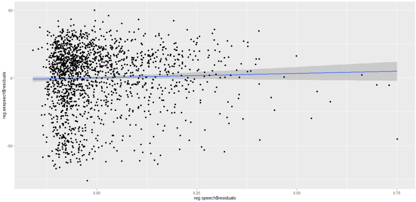


Loudness

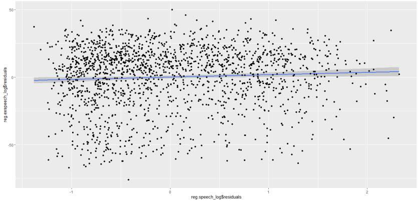
Linear Regression



Mode

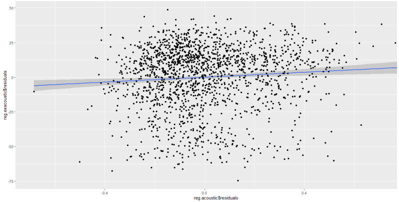


Speechiness

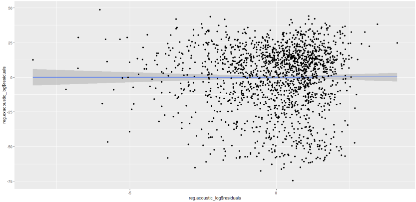


log(speechiness)

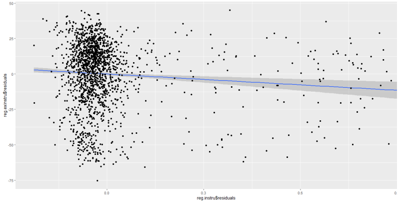
Linear Regression



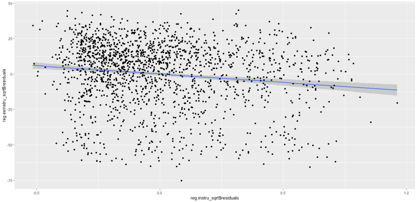
Aconsticness



log(Aconsticness)

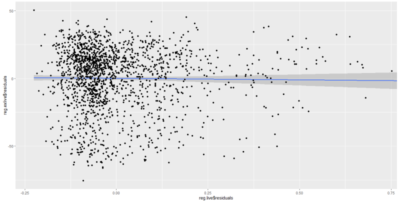


Instrumentalness

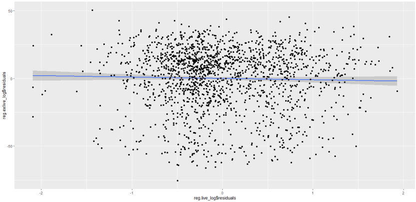


Sqrt(Instrumentalness)

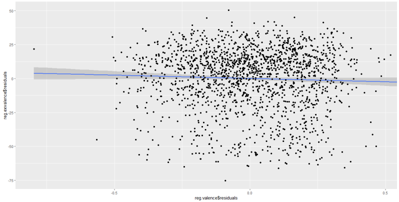
Linear Regression



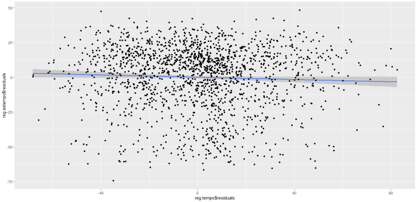
Liveness



log(Liveness)



Valence



Tempo

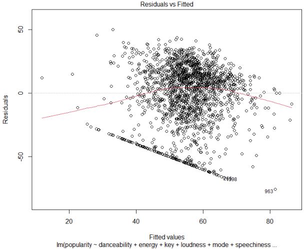
Linear Regression

From the partial coefficients plots above, we have the following conclusions:

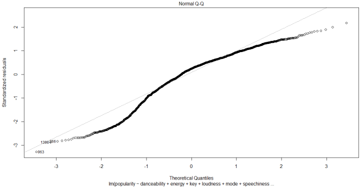
1. There isn't much linear relationship between some explanatory variables and response variable.

2. Some transformations (such as log or sqrt) of explanatory variables can better the regression effect.

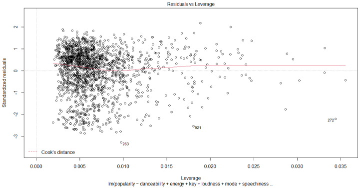
Linear Regression



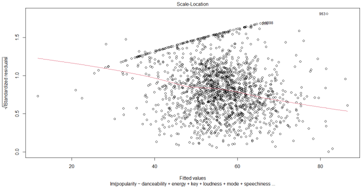
Residual vs Fitted



Normal Q-Q



Residuals vs Leverage



Scale-Location

Linear Regression

From the diagnostics above, it's obvious that the variance is unequal and abnormal, violating the assumptions.
Also, there are outliers with high leverage, lowering regression effect.
We further use BoxCox to find out the suitable transformation, as well as the introduction of interaction terms.

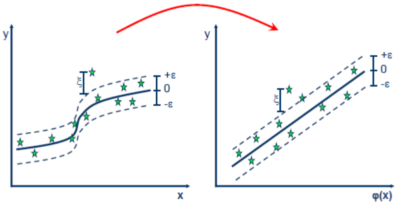
1. Exploratory Data Analysis
2. Model selection
   - Linear model
   - SVM regression
   - Tree method
3. Some questions

SVM regression

As we have found that linear model really couldn't give a convincing result, we tried support vector machine with kernel trick, mapping our low-dimensional data to a high-dimensional space to make it easier to fit a linear model.

Pros

- Compared with linear model, SVM will penalize only those $\hat{y}$ that lies out of the "margin", therefore may tolerate noise more and should be more suitable when dealing with data from real world.
- Kernel trick helps us to automatically add some non-linear terms, and we don't need to find some specific non-linear terms to improve our model manually.



Cons

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (\alpha_i - \hat{\alpha}_i) K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

- Kernal trick makes the model like a "black box", and the result is not that interpretable. (especially when we use RBF kernel, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}{2\sigma^2}\right)$)
- MSE on test data: 612.4

Tree method

- We will try GBRT (gradient boosting regression tree) later on.
- Good at handling different types of data
- Strong predictive ability
- Robust to outliers

Some questions

- How can we improve our linear model?
- Should we concentrate on one model or retain the three models?
- ...

Thanks!