UNIVERSITY OF WISCONSIN-MADISON

KKBOX Music Database

Exploratory Data Analysis

ZHANG Congfei
ZHU Wenjun
2017-12-7

1. Introduction

As we enter the 21th Century, there is a boom in the use of electronic device to enjoy music, taking the place of traditional music playing devices such as radios. Moreover, with the development of Internet, nowadays there are a great number of online music players, which enable us to record and track favor of users. KKBOX is Asia's leading music streaming service and holds a music library of over 30 million tracks.

In this project, we are interested in getting a better understanding of the users' features, the features of songs and how the features combine together to determine how likely a song is being liked.

2. Data Processing

2.1 Dataset

The data is acquired from the website of Kaggle[1] and contains 4 files called train, songs, members and song_extra_info respectively. The dataset documents information about different songs and different information about users, including user ID, user gender, song ID, song length, song name, artists, target, etc.

Important variable:

Target: This variable denotes whether or not the user replays a song after him/her first get exposed to the song. (i.e. the user likes the song.) This variable is ultra-important because we use this variable to calculate a lot of like rates based on it.

2.2 Data Preprocessing

We have 4 data sets with different set of data, in order to obtain a consist data form, we first use the inner join function provided by pandas in python to join the data sets together.

After the join, for each single row of data, we include information as follows (20 Columns in total):

Msno (User ID)	song_id (Song ID)	source_system_tab (Source System)	source_screen_name (Access method)	source_type (Source Type of a song)
Target	Name (Name of Song)	Isrc (Id of lyrc)	City (City of a User)	Bd (Age of User)
gender	registered_via (Registration Method)	registration_init_time (Registration Time)	expiration_date	Language (Language of a song)
song_length	genre_ids (Genre id of a song)	artist_name (Name of Artist)	Composer (Name of Composer)	lyricist

You may have realized that there should be duplicate values of user's ids in the data base. The answer is yes. Each row describes an **event** in brief: a certain song is recommended to a certain user, and the user like the song or not, together with additional information about the user and the song.

As what's usually happening, there exists a lot of null values and outliers that is not correct in the data set.

In order to do further analysis to the data set, we clean the data first.

There are lot of dirty data, some we can fix, and some we cannot. For the genre column, they are integers, but we can't find the corresponding list of name of genres, thus we can only discard the column.

For the language column, it's also in integers. Lucky, for this feature, we can inspect the corresponding name of song and determining the actual corresponding language to each integer.

For the bd (language) column, there are a lot of impossible outliers in this feature (e.g. -10, 900, etc..) we also don't have the technic to verify the real value of the outliers. Thus when doing analysis related to age, we discard the outliers.

For the composer& artists & lyricist columns, there exits different names but definitely pointing to the same composer& artists & lyricist. We group the data sets by grouping manually, pattern matching, and similarity calculation to merge them into one single value.

For the city, again it's not meaningful for the EDA analysis, we just discard them now. Maybe we can come back and see if we can get the geometric information corresponding to each integer in this column.

Similar operations to features containing integer values and not have a meaningful name string.

The cleaning is mostly done in python by certain algorithms, some of the manual operations are done in tableau.

After the preprocessing, we finally can start the analysis.

3. Data Analysis

Recall the objective of our project, the main object of our project is the users, there is no doubt we should start our exploration with the users.

3.1 User Analysis

We first plot the registration number with respect to different years.

In order to get detailed information about the ages of users, we generate a shadow plot with respect to ages. In order to make the graph consist and meaningful, we first group the ages into different groups, with a width of 10 in each bin (without the outliers).

With the proliferation of online music services, the registration number of KKBOX

has gone through a dramatic increase over the years. As can be seen from Figure 3.1.1 below, the total registration number in 2016, regardless of different age groups, is almost eight times the original amount in 2004.

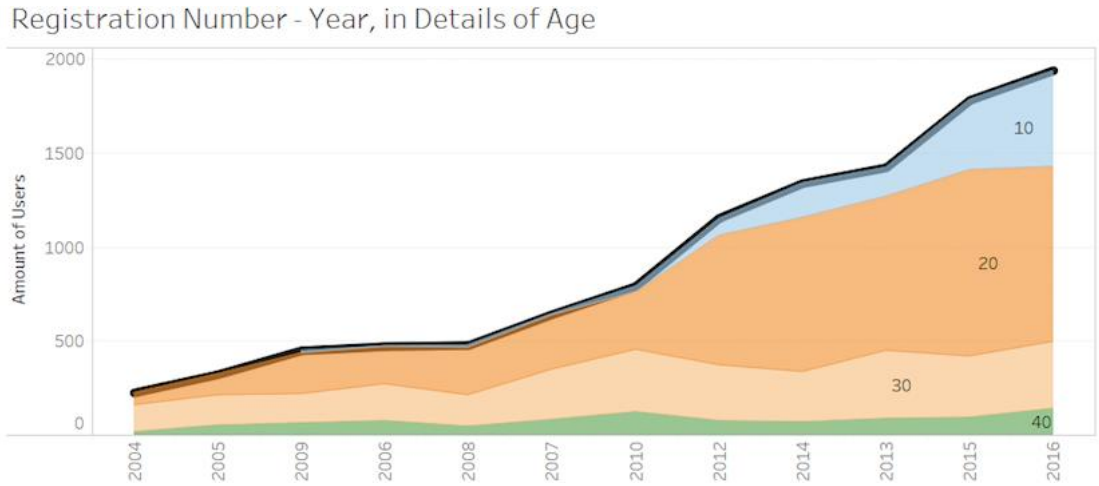

Figure 3.1.1: Registration Number

We see a trend in the percentage of age groups in the database, this is interesting. It's useful to find the construction of age groups in the database so that we can choose to bias the most potential groups in advance to grasp the market. The Result is Shown in Figure 3.1.2.

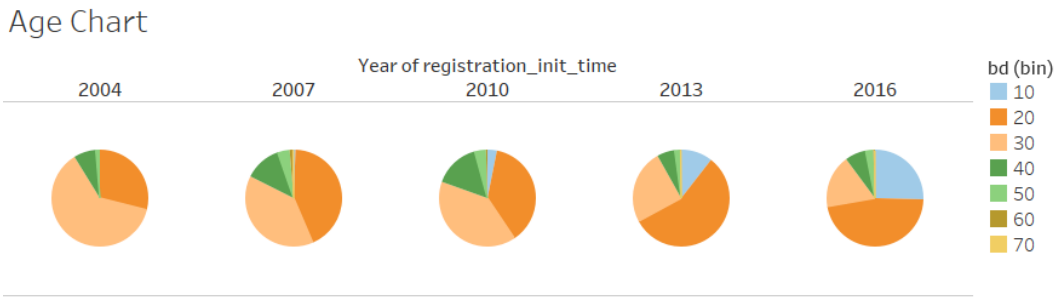

Figure 3.1.2: Age Distribution

As to age distribution, from the pie chart in Figure 3.1.2, there are noticeable changes on the proportion of different age groups over the years. Combining the result from Figure 1 above, we can see that there are significant increases in the registration number in the age groups of 10~20 and 20~30, while the registration number in the age groups of 30~40 and 40~50 increased only a small amount. The proportion of age group 20~30 has overtaken age group 30~40 to be the most outstanding part of age chart. And age group 10~20 is obviously a potential stock.

After discovering the age distribution, we are also interested in the gender distribution in the users. From Figure 3.1.3 below, we get the amount of male users and female users respectively. They all keep an overall increasing trend over the years. Utilizing the FORECAST function of tableau, we get the predicted amount of registered users in year 2016 and 2017 in both genders separately.

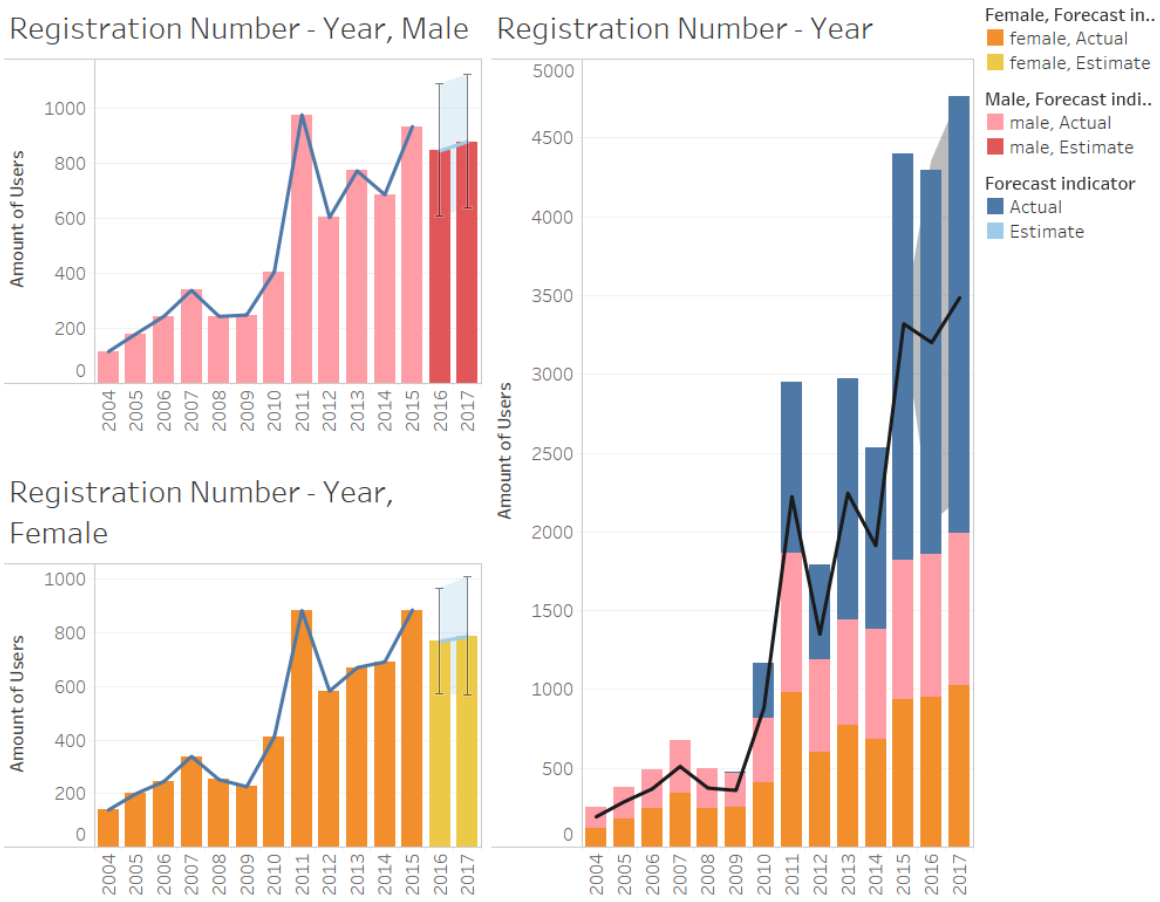

Figure 3.1.3 Gender of Users

From the graph we can see if we just discard the users with null genders, we get almost the same number of user in male and female. Which means the users in KKBOX doesn't have a bias towards gender, i.e. we don't have a certain gender of users that loves to use KKBOX as their music player more.

How do users usually get exposed to new a song?

To answer this question, we plot the count of users getting exposed & listening to a song.

Figure 3.1.4 shows different methods for users to get exposed to songs, containing count of songs and replay rate for different methods. Among those online methods, Discover is the highest in respect to replay rate, with Album comes second and then Artist. So it is feasible for KKBOX to highlight the part of Discover to attract users to listen to new songs, and then to add it into their local list and replay it later.

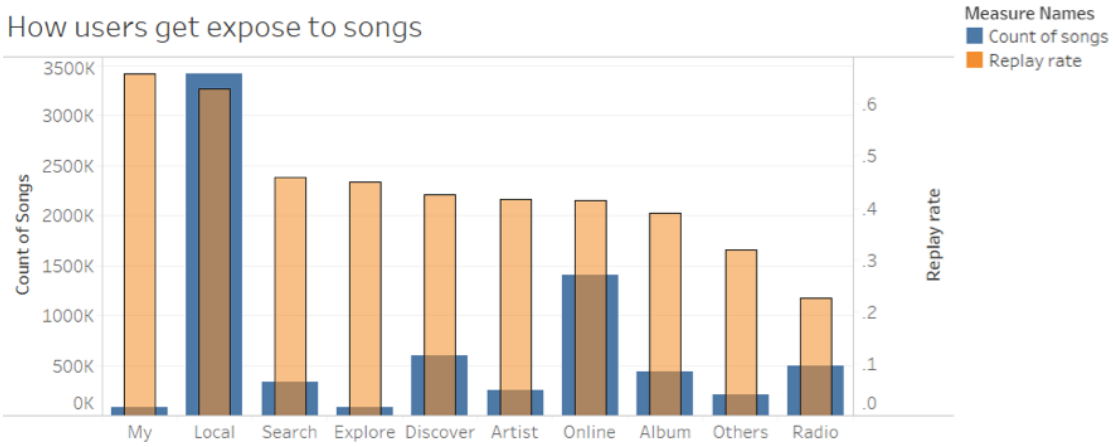

Figure 3.1.4: Method to Get Exposed to Songs

Figure 3.1.5 shows different methods for users to listen to songs. We can see that users usually choose to listen to local songs than to listen to new songs online, and thus replay rate of local songs and those in "My library" are much higher than other methods. And again Discover ranks first among the online methods. So we get the same result.

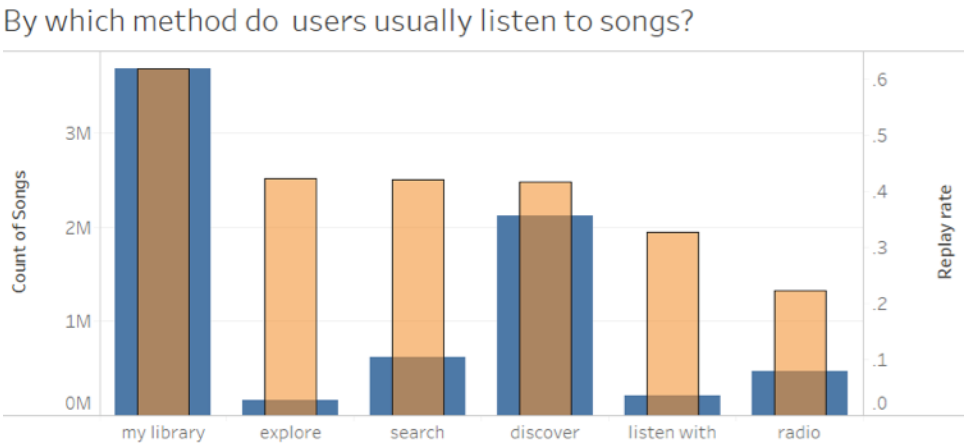

Figure 3.1.5: Method to listen to songs

3.2 Song Feature Analysis

Among the database, which language of songs dominates the database (Most being liked)?

To answer the question, we first analysis the language of songs appeared in the database.

There are about 20 languages in total, and we only care about the most popular languages, thus we plot the pie chart and remain only the 4 most popular languages and group the other languages as "Others" in the chart. The result is shown in Figure 3.2.1.

Define the angle of the pie chart for each language as:

$$\frac{(sum\ of\ targets\ grouped\ by\ language)}{total\ sum\ of\ targets\ in\ the\ database}$$

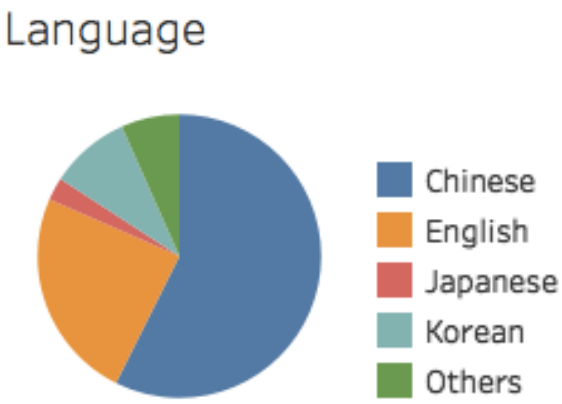

Figure 3.2.1. Language Distribution

From Figure 3.2.1, we can see that Chinese, English, Japanese and Korean are the four most popular languages on KKBOX, with Chinese acts as the most outstanding one. It can be understood as KKBOX is an Asian music service.

As a result of language analysis, we show our interests in songs in the 4 languages.

We want to find the best Artists and Composers in different languages, so we plot the rankings of the most popular Artists and Composers in different languages.

We rank the singers by calculating **sum of target over all his/her songs** and use different color (blue to red) to represent replay rate. As a result, we get the most popular singers in the four different languages respectively in Figure 3.2.2 below. There is no wonder that Jay Chou comes out at the top of the Chinese list, with a fairly high replay rate as well. He has always been regarded as Asian Pop King. Japanese rock band RADWIMPS, American rock band Maroon 5 and Korean band BIGBANG ranks first on Japanese list, English list and Korean list respectively. We can also see that click rate and replay rate of Chinese singers are comparatively high, while both rates of Japanese singers are comparatively low.

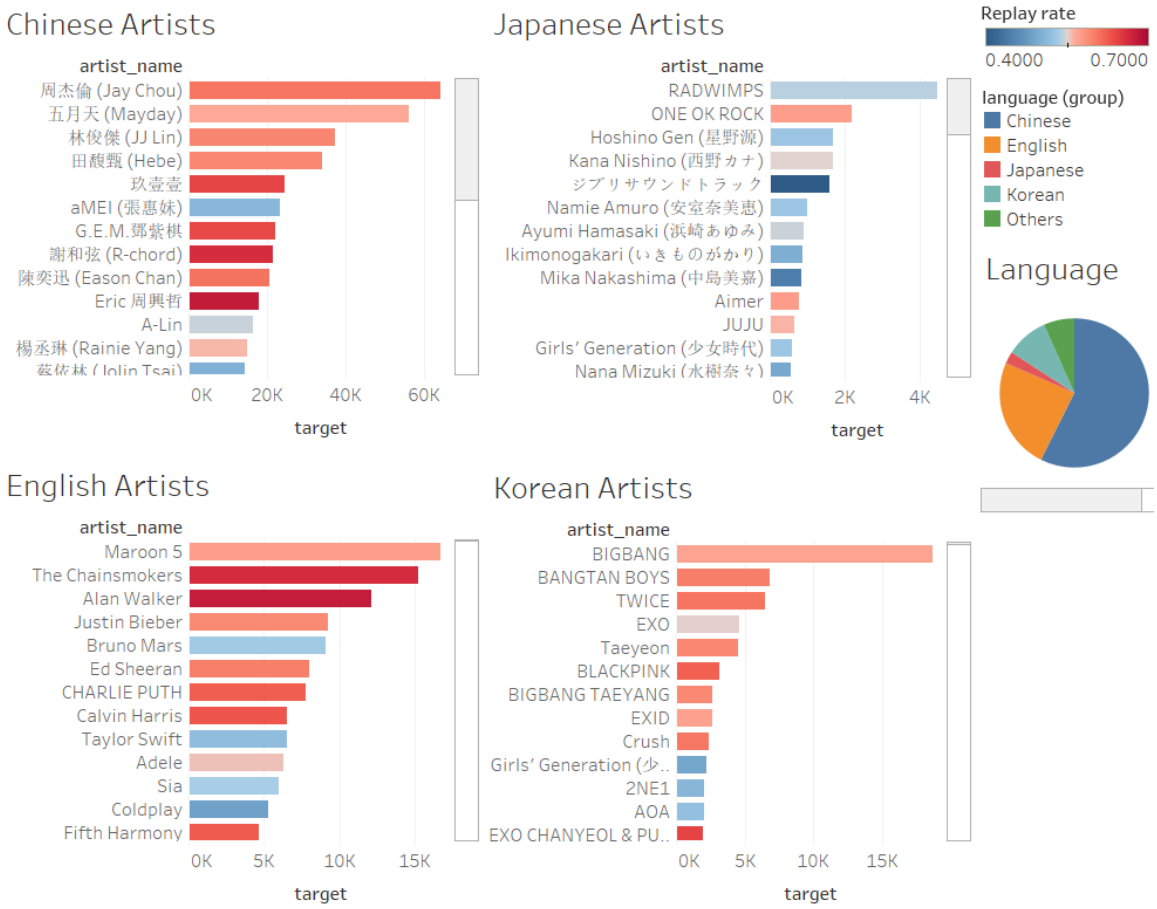

Figure 3.2.2: Most Popular Artists

Similarly, we get the most popular composers in Figure 3.2.3 below.

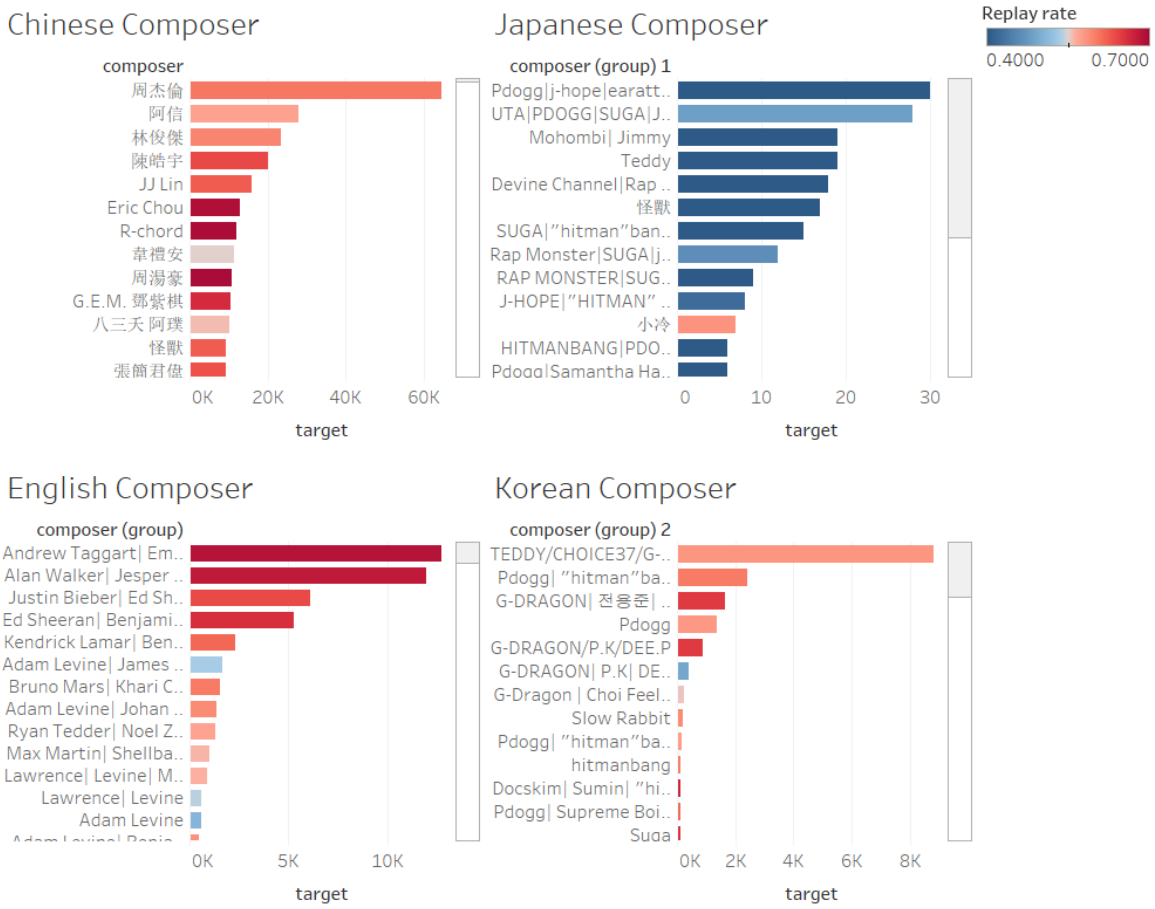

Figure 3.2.3: Most popular Composers

By the previous analysis, we get the most popular Artist and Composers, but what are the most potential ones?

We define the potential stars as the Artists whose sum of target over all his/her song is more than 1000, and less than 5000. We rank the potential stars with its average replay rate. We define the average replay rate as:

$$\frac{sum\ of\ targets\ grouped\ by\ Artist}{sum\ of\ count\ of\ targets}$$

This variable denotes the probability of the artist's song being liked when being recommended to users, which can be an indicator or the quality of the artists' works.

Then we group the potential stars by their language, we can see potential start comes most from Chinese language.

We plot the result of this analysis in Figure 3.2.4.

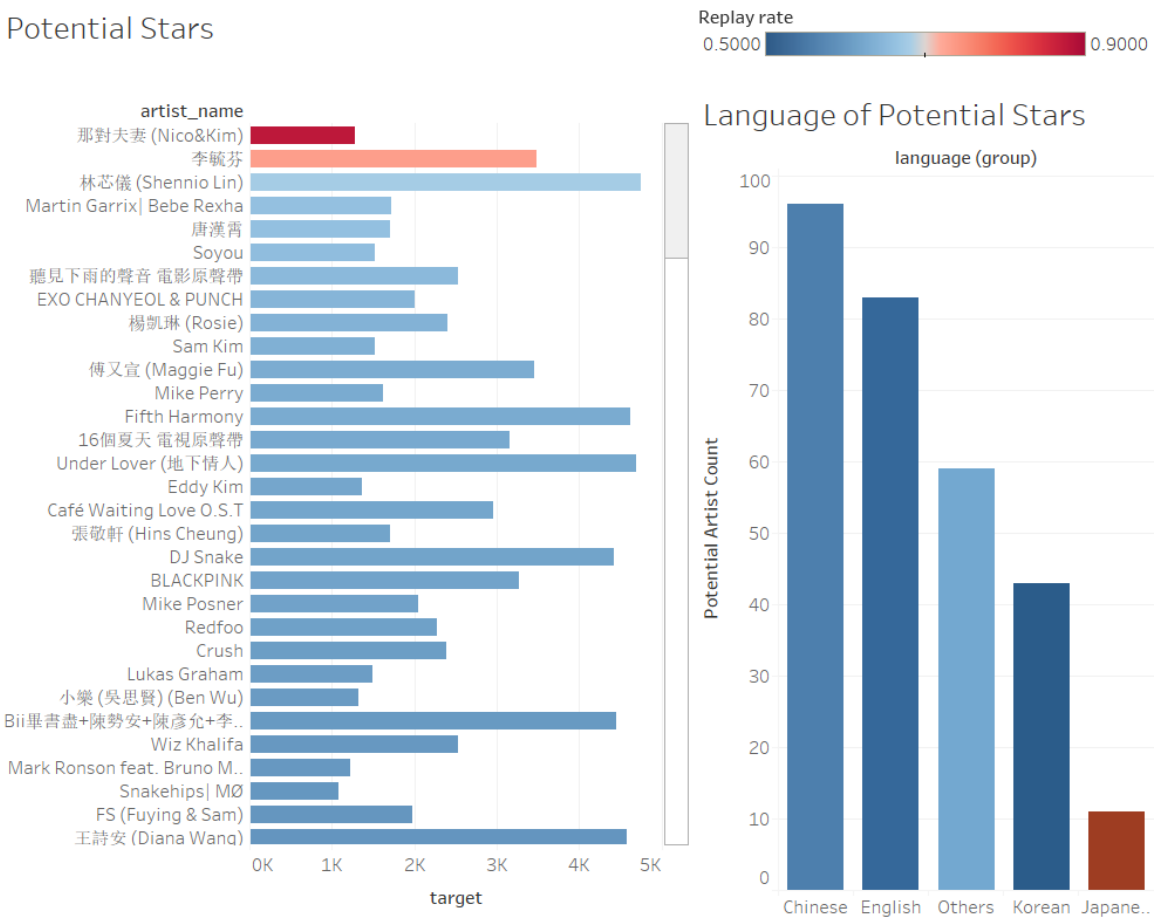

Figure 3.2.4 Potential Stars

3.3 Song Length Analysis

We are interested in finding the relationship between song length and replay rate.

To start with, we first create bins with a length of 30 seconds and group the songs into bins. Then we generate figure 3.3.1.

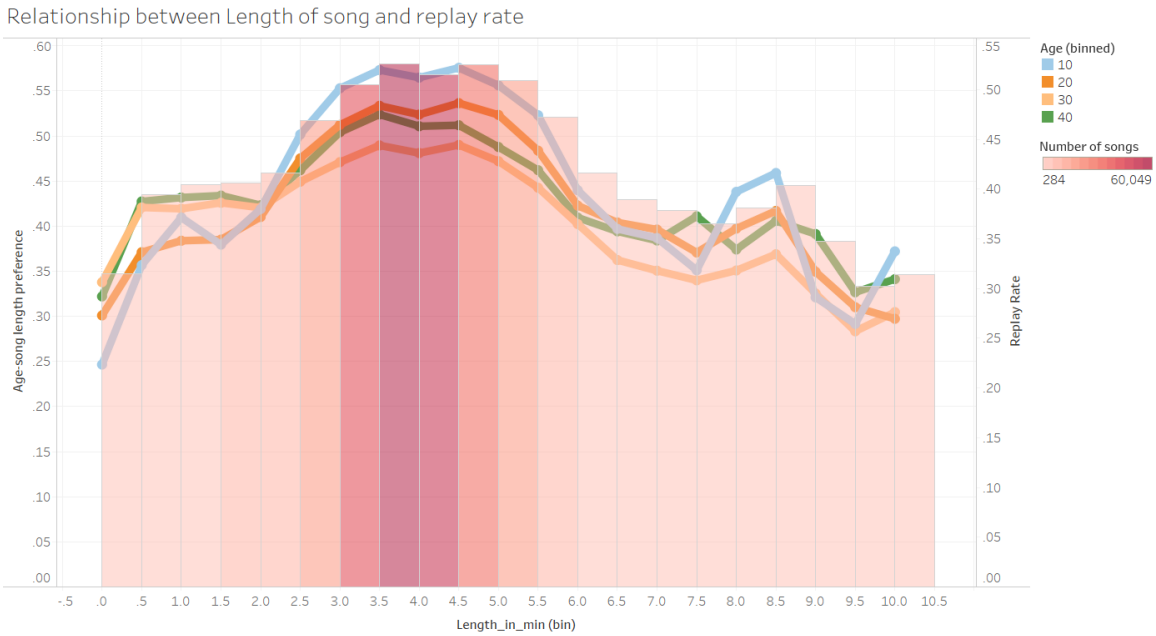

Figure 3.3.1 Song Length Analysis 1

Again, we plot the like rate with respect to age.

The height color bars show the average replay rate inside the category of songs and the color of color bars indicates how much songs are classified in this category.

Surprisingly, we do find there exists a relationship between the length of a song and the like rate. The chi-squared test also confirmed our observation from the graph.

It seems that users aged from 10-20 is the group like songs with a length from 3 min to 5.5 min the most from the graph.

But wait, is this the real case?

The answer is **NO**!

This is not fair for the other age groups; they may have a lower average like rate. So it's not fair to compare only the like among the songs when we are thinking about the bias of users to songs.

In order to get a fair measure, we divide the like rate of a certain age group towards songs in a certain length by the average like rate of users in the certain age group.

The higher the rate is, the more biased the group of user is to the songs.

Thus we get figure 3.3.2

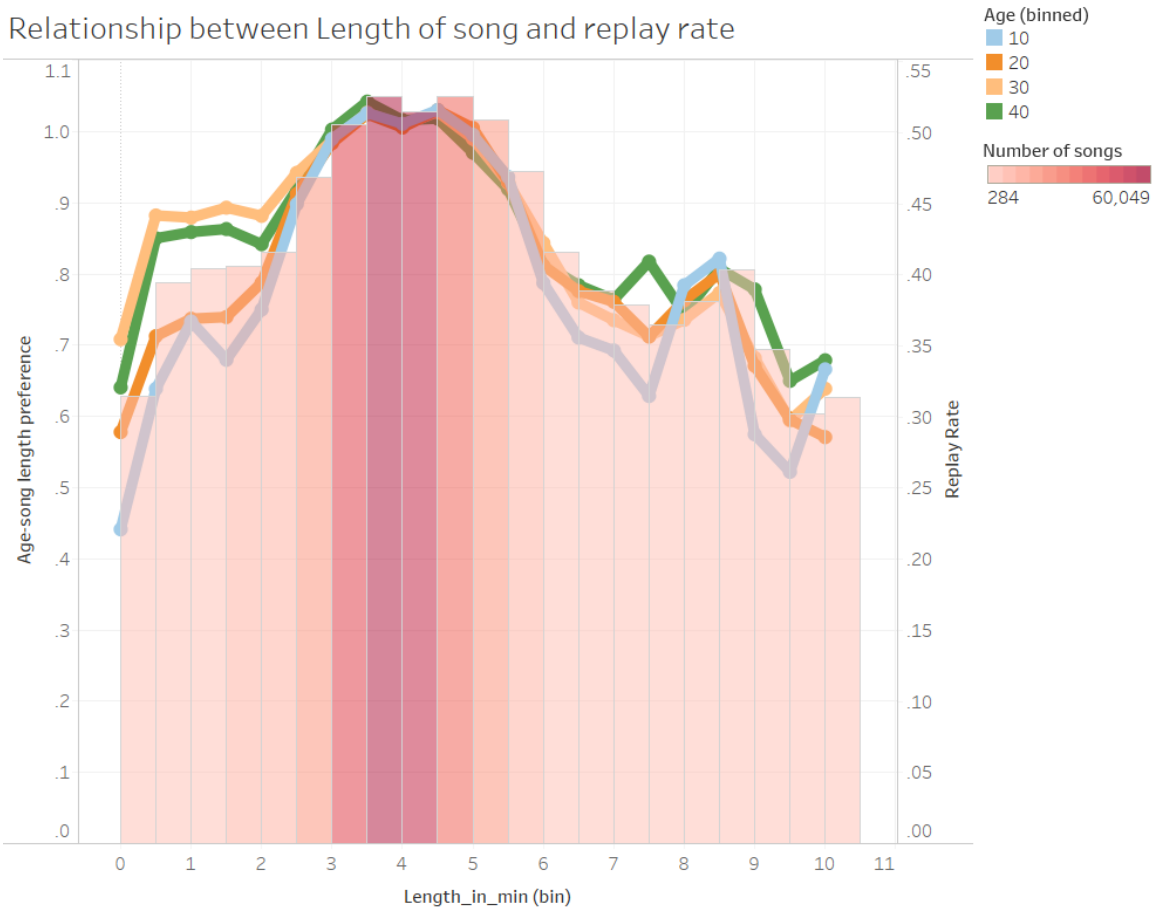

Figure 3.3.2 Song Length Analysis 2

The Result is interesting: All the group of users are biased to the songs with a length from 3 min to 5.5 min, and the users over 30 like songs with a shorter length much more than the younger users.

3.4 Relationship Analysis

In this section, we try to find what kind of cooperation between Artists, composers and Lyrists gives a boost to the Artist.

We use heat maps to show the results (Figure 3.4.1).

In each square, we use the size to the popularity of the artist and the color to denote the replay rate. The deeper the color is; the largest replay rate we have.

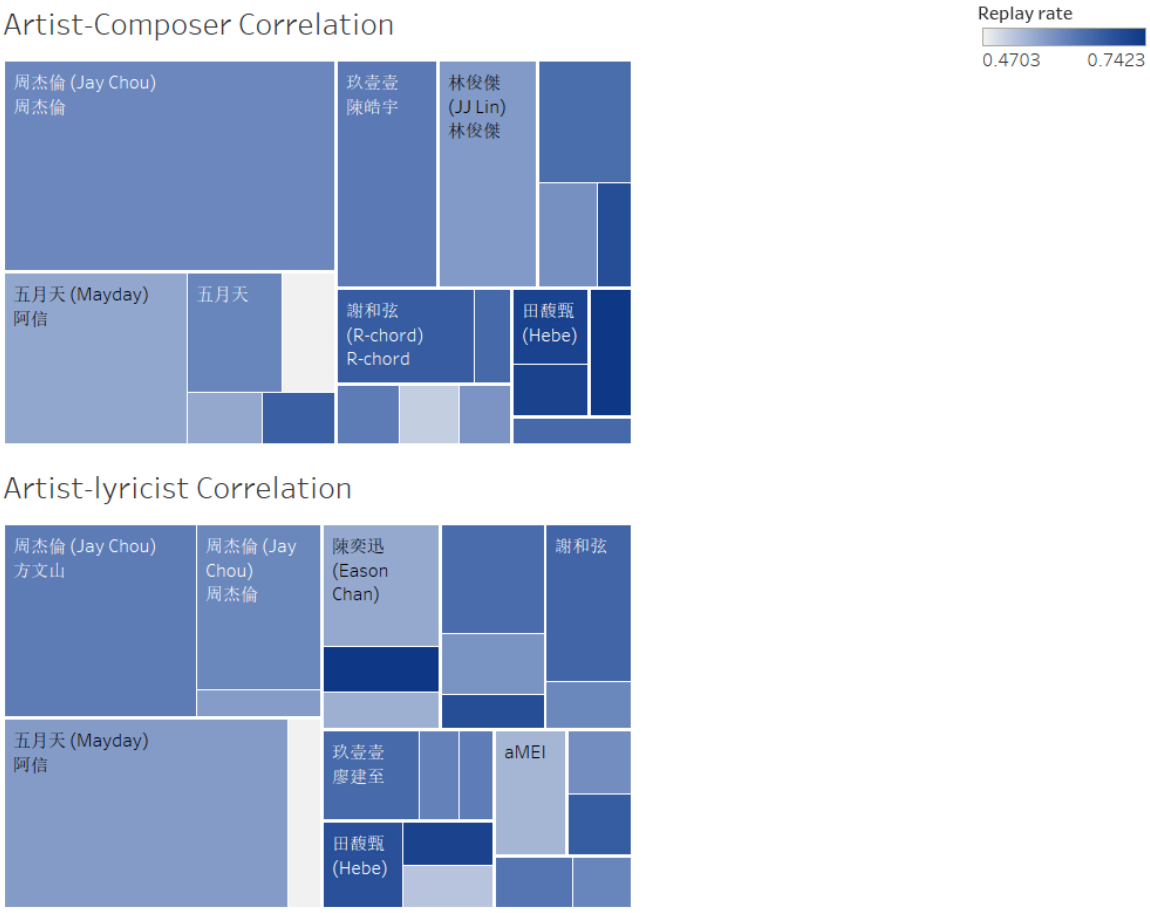

Figure 3.4.1 Cooperation 1

Wait, we quickly realized that this is also biased. For talented Artists, they usually have a higher average replay rate, and sometimes it's the dominant factor.

Thus in order to find a proper indicator to show the boost given by cooperations, we divide the replay rate of the cooperation by the average replay rate of the artist to find the results again. The results are shown in figure 3.4.2.

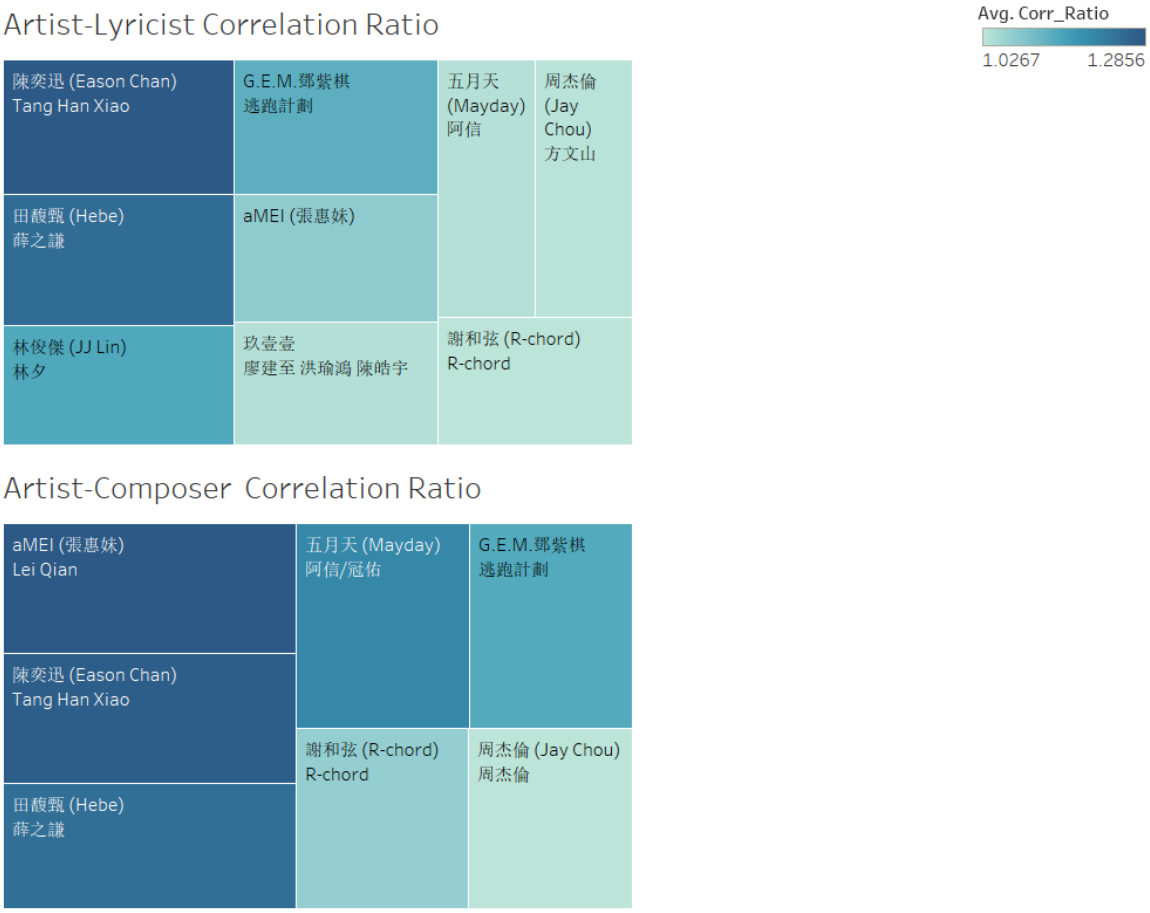

Figure 3.4.2 Cooperation 2

We find the results changed significantly, the factors of artists itself IS very likely a dominant factor in the replay rate.

4. Conclusion

4.1 Conclusion

In conclusion, the registration number of KKBOX users is growing steadily over the years, with the age group of 10~30 make up the most significant part. So they are the target customer group, which KKBOX should take most effort to attract.

We find out that the most popular song language is Chinese, together with English, Japanese and Korean. We also find out the most popular singers and composers in the four languages. Moreover, we discover the best cooperation between singers and composers, as well as lyricists.

Length of songs does affect the probability a song being repeated and 3~5 minutes-long songs are more likely to be repeated. Though users listen to songs in their library the most, when they listen to songs online, they are more likely to replay a new song found in the Discovery part. So we recommend KKBOX to highlight the part of Discover to attract users.

5. References:

[1]. https://www.kkbox.com

[2]. https://www.kaggle.com/c/kkbox-music-recommendation-challenge/data