KKBOX Music

Congfei Zhang & Wenjun Zhu

# 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, 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 determine how likely a song is being replayed.

# Analysis

## Dataset description

The data is acquired from the website of Kaggle. The dataset documents information about different songs and users, including user ID, user gender, song ID, song length, song name, artists, target, etc. Here target is a variable that denotes whether the user replays a song after him/her first get exposed to the song. This variable is very important as we use it to calculate “like rates”.

## Software used

We used Python to preprocess the data, rejecting entries containing NA (missing value); then used Tableau 10.5 to visualize the data.

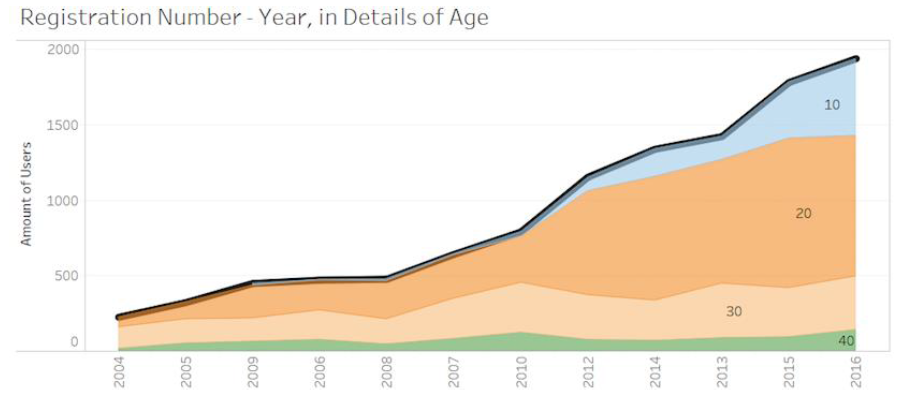
## Approaches

### User analysis

We first plot the registration number in different years. To get detailed information about the ages of users, we generate a shadow plot according to ages. We divide the ages into different groups, with a bin width of 10.

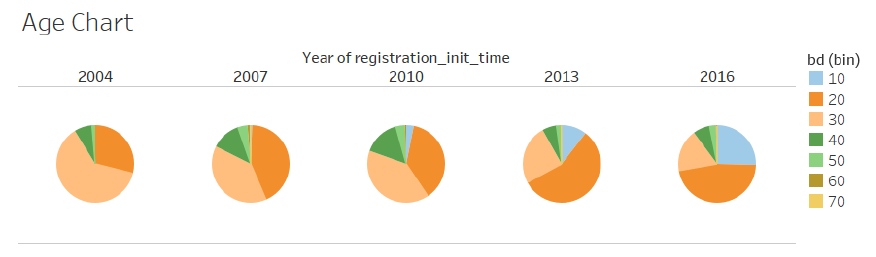
As can be seen from Figure 1 below, with the proliferation of online music services, the registration number of KKBOX has gone through dramatic increase over the years. The total registration number in 2016, regardless of different age groups, is almost eight times the original amount in 2004.

Figure 1. Registration Number



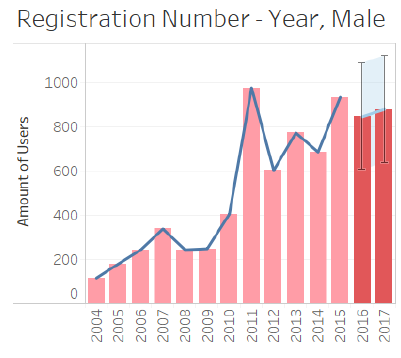
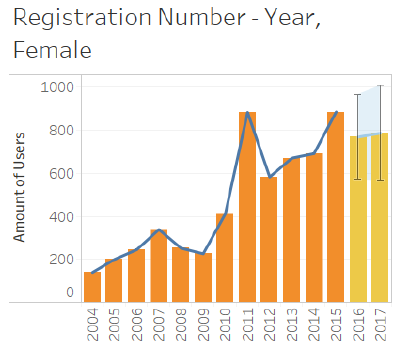
It’s useful to analyze the composition of age groups in the database. we can choose the most potential groups of new consumers in advance to grab market share. From the pie chart in Figure 2, there are noticeable changes in the proportion of different age groups over the years. Combining with Figure 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.

Figure 2. Age Composition



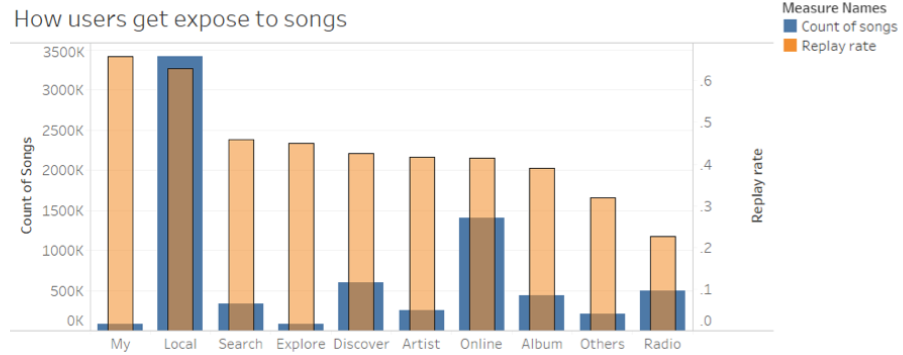
Apart from age distribution, we are also interested in the gender distribution. From Figure 3 below, we get the number of male users and female users respectively. They all keep an overall increasing trend over the years. Utilizing the FORECAST function of tableau, we get the predicted number of newly registered users in year 2016 and 2017 in both genders separately.

Figure 3. Gender Composition

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 4 shows different methods for users to get exposed to songs, containing count of songs and replay rate. Among all the online methods (apart from self-search and explore), Discover is the highest in 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 song.

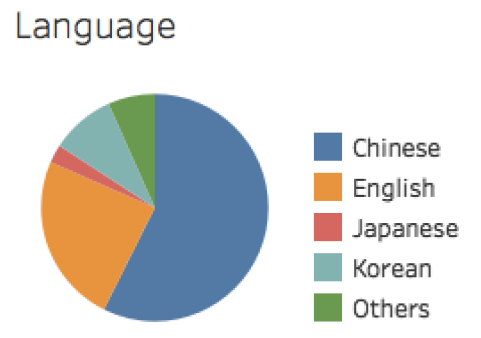
Figure 4. How do users get exposed to songs?



### Song feature analysis

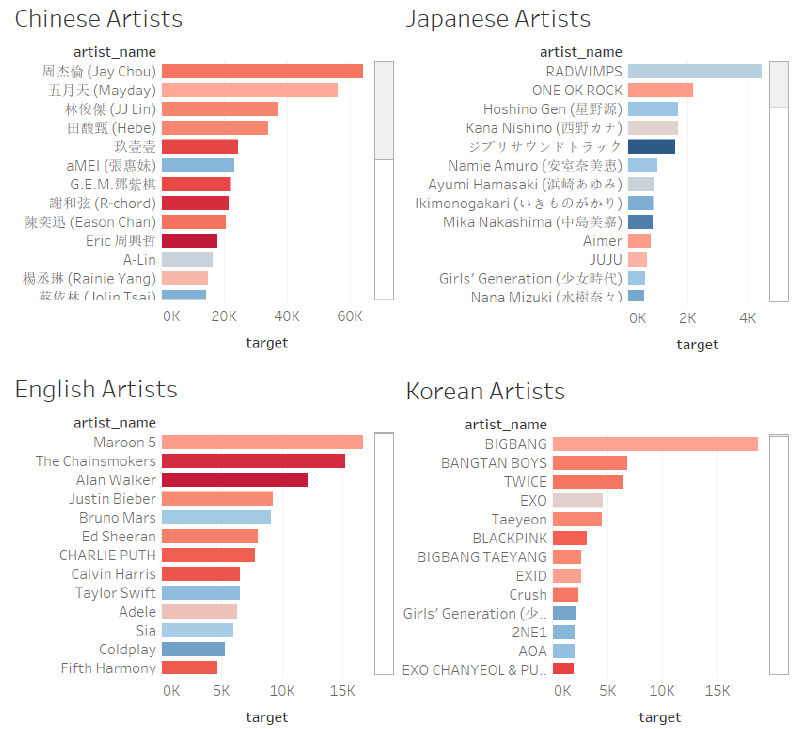
Among the database, which language of songs dominates the database (Most being liked)? There are about 20 languages in total, and we only care about the most popular languages, thus we plot the pie chart and remain only the 4 most popular languages and group the other languages as “Others” in the chart. The result is shown in Figure 5. The result is no wonder, as KKBOX is an Asian music service.

Figure 5. Language proportion



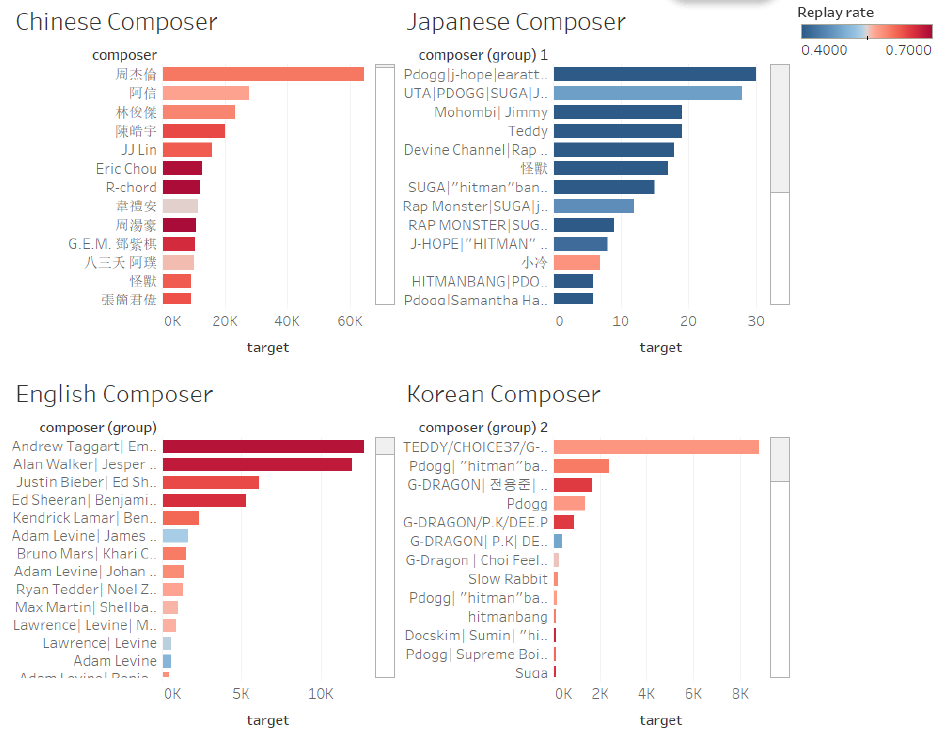
We then want to find the most popular Artists and Composers in different languages. So, we plot the rankings of Artists and Composers in different languages. We rank the singers by calculating sum of target over all his/her songs and use different color (blue to red) to represent replay rate. We get the most popular singers in the four different languages respectively in Figure 6 below. There is no wonder that Jay Chou comes out at the top of the Chinese list, with a high replay rate as well. He has always been regarded as Asian Pop King.

Figure 6. Most popular artists



Similarly, we get the most popular composers in Figure 7 below.

Figure 7. Most popular composers



We then 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. In each square, we use size to represent the popularity of the artist and color to denote the replay rate. The deeper the color is, the larger replay rate we have. From Figure 8 and Figure 9 below, we can find the boost caused by corporation.

Figure 8. Artist-composer relationship

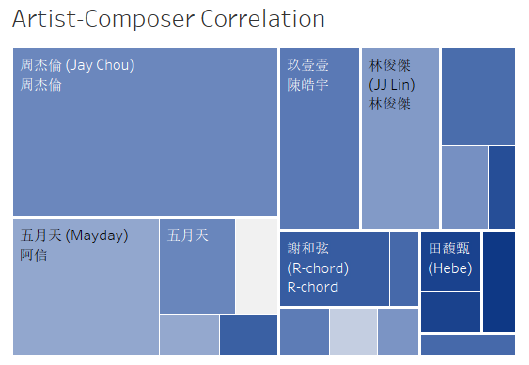
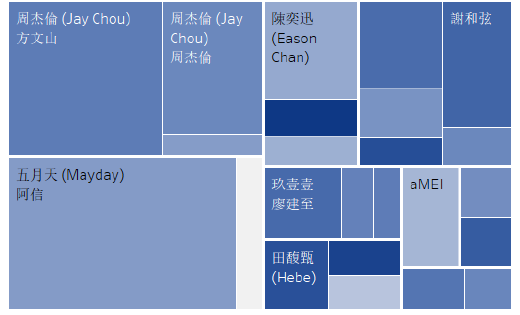
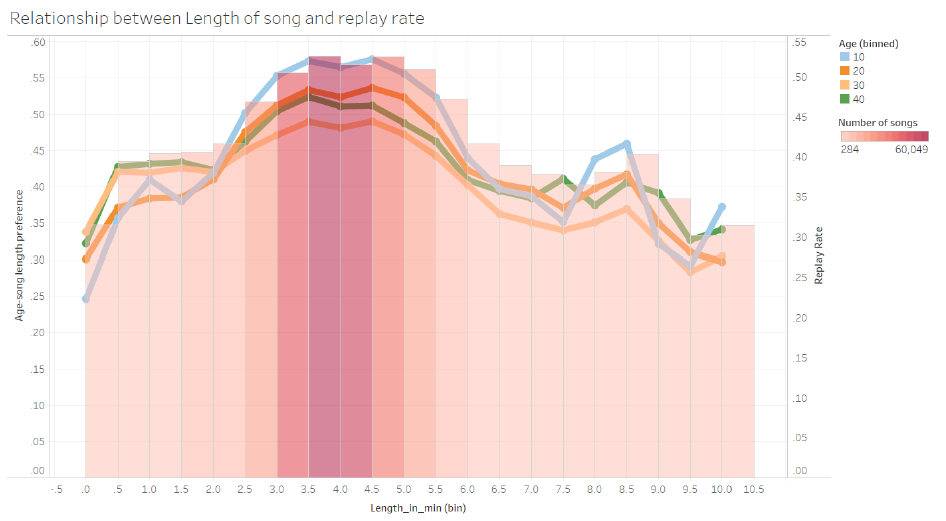


Figure 9. Artist-lyrist relationship



We are also interested in 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 10.

Figure 10. Relationship between song length and replay rate



Again, we plot the like rate with respect to age. The height of bars shows the average replay rate inside the category of songs; and the color of bars indicates how much songs are classified in this category. We do find there exists a relationship between the length of song and the like rate. It seems that songs with a length from 3 min to 5.5 min are most likely to be replayed.

# 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.